SYNTHESIZING DATA SOURCES TO DEVELOP AND UPDATE RISK MODELS

by

Paige Maas

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

June, 2014

Abstract

Building risk models from multiple different sources of data allows researchers to incorporate the best available information on key model parameters. In this thesis, we develop and apply methodology for optimally combining information from multiple data sources in two main contexts.

In the first, motivated by the need for building subtype-specific absolute risk models for breast cancer, we develop and apply methodology for combining information information from analytic cohort or case-control studies and from population-based registries. We address the statistical challenges involved with handling different types of missing information in this context. We derive variance estimators for the risk predictions produced by such models, accounting for different sources of uncertainty. We apply the methods to two large consortia in order to build absolute risk models for overall breast cancer and for subtypes of breast cancer defined by estrogen receptor status. We show how the absolute risk models can be used to project distributions of breast cancer risk for the US population and to evaluate the potential impact of population-wide modification of breast cancer risk factors.

In the second problem, we consider the issue of how to effectively incorporate external information when building new or updated risk models, again with the goal of combining data sources to produce models that are more efficient and representative of the underlying population. In particular, we explore a regression calibration approach, utilizing a method from sample-survey literature which is traditionally used for increasing the efficiency of parameter estimation from a given survey by leveraging information from external data sources. We examine the performance of the estimator in a context that has not previously been studied, where the sample and the

external data are representative of different populations. We derive theoretical conditions under which the calibrated estimator produces meaningful estimates, which are calibrated to the external population, and corroborate our analytic results with numerical simulations. Our work also identified weaknesses in the methodology and promising avenues of further research in this important area.

Thesis Readers:

Nilanjan Chatterjee, Ph.D. (advisor)

Mei-Cheng Wang, Ph.D. (co-advisor)

Xiaobin Wang, M.D. (committee chair)

Kala Visvanathan, M.B.B.S.

Ravi Varadhan, Ph.D. (alternate)

Alden Gross, Ph.D. (alternate)

Acknowledgments

I am grateful to the many talented and caring people who have inspired and supported me throughout my graduate training. I am incredibly fortunate to have worked with my advisor Nilanjan Chatterjee from whom I have learned a tremendous amount. I am especially thankful for the opportunities he provided me to work directly with exceptional datasets. I am also grateful to my advisor Mei-Cheng Wang for her guidance in the thesis writing process and in my personal transition from Hopkins to the National Cancer Institute. I want to sincerely thank Alden Gross, Ravi Varadhan, Kala Visvanathan, and Xiaobin Wang for taking the time to serve on my thesis defense committee and for providing valuable, constructive feedback.

Thanks to the Biostatistics in Cancer Epidemiology Predoctoral Training Program, I have had the privilege to work with and alongside many committed researchers in the Division of Cancer Epidemiology and Genetic's Biostatistics Branch (BB). By sharing their insights in discussions at journal club, tea time, and seminar, the members of BB have had a significant impact on how I think about public health research. I am grateful to my collaborators Bill Anderson, Raymond Carroll, Roni Falk, Mitch Gail, Cathy Schairer, and Regina Ziegler for sharing their expertise with me and treating me as a valued member of the collaborative team. I would like to give a special thanks to my collaborator Montse Garcia-Closas for sharing her knowledge in the area of breast cancer risk prediction and for welcoming me into her home. I am also especially grateful to postdoctoral fellows Stephanie Kovalchik and Simina Boca for their mentorship during our shared time in BB.

I am grateful to have been a part of the vibrant academic community in the Biostatistics Department at Johns Hopkins. I am especially appreciative of the department's continued support after my transition to BB. I am fortunate to have been

able to learn from a remarkable group of people, including both my professors and my peers. I note in particular Francis Abreu, Alyssa Frazee, Sarah Khasawinah, Jenna Krall, and Yenny Webb-Vargas with whom I've shared countless moments of commiseration and celebration. I am constantly inspired by the many ways they exemplify perseverance with a positive attitude, and I am lucky to have counted them as friends throughout my graduate program. I also want to thank my undergraduate professor Jo Hardin, whose teaching and encouragement inspired me to pursue statistics in graduate school.

I am thankful for the support of my family, both in graduate school and beyond. I thank my parents, Tracy Oliver & David Maas, and my grandparents, Donald & Gloria Oliver and Mary Lynn & David Maas, for providing financial support for my education, not to mention all of my genes. The encouraging words of my grandparents Donald & Gloria Oliver, and their constant support and interest in my work throughout graduate school, has meant a lot to me. I am grateful to my sister Casey for her upbeat and unwavering confidence in me and for providing opportunities to practice my statistical consulting skills. Last, but by no means least, I thank Daniel Oppenheimer for the uncountably many ways he has made my life better throughout this challenging process.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Absolute risk models predict disease risk in an upcoming time interval based on known risk factors for an individual or individuals in a population, accounting for the presence of competing risks (Gail et al., 1989). Absolute risk models for cancers and other diseases have important clinical and public health applications.

Absolute risk models can be used to identify individuals at high risk of disease in order to target screening and disease prevention strategies (Jackson, 2000; Jackson et al., 2005; Pharoah et al., 2008; Gail, 2011). In the past, decisions regarding the initiation of screening or preventative intervention have often been made on the basis of age and family history, as proxies for risk. However, there is increasing consensus in the medical community that these decisions should instead be guided directly by individualized estimates of risk, which can be obtained from absolute risk models that include a wider array of environmental and genetic risk factors.

At the public health level, direct estimates of risk allow researchers to quantitatively weigh the risks and benefits of a particular screening regime or preventative intervention and tailor those strategies in a way that is optimal for the underlying population (Grundy, 1999; Gail, 2001; Murray et al., 2003). An example of this in

practice is the American Society for Colposcopy and Cervical Pathology Consensus Guidelines for cervical cancer screening, which are based on quantitative evaluation of the benefits and potential harms of screening as measured by absolute risk (Saslow et al., 2012). Other examples include using absolute risk models to identify absolute risk thresholds for which the benefits associated with preventative breast cancer treatments, such as Tamoxifen, outweigh the risks associated with treatment (Chlebowski et al., 2002) and to evaluate the impact of smoking cessation on lung cancer (Halpern et al., 1993). Absolute risk models can also be used to determine the necessary sample size for prevention trials by projecting the expected distribution of disease risk based on the distribution of risk factors in a population (Gail, 2011).

At the patient level, absolute risk estimates can be used to counsel individuals on the basis of their personal risk. In fact, the National Cancer Institute has created a number of risk assessment tools for this purpose, which are available online. The Breast Cancer Risk Assessment Tool estimates a woman's risk of invasive breast cancer based on responses to 8 questions about her age, race, and reproductive and medical history (Gail et al., 1989). The Colorectal Cancer Risk Assessment Tool estimates risk of colorectal cancer for individuals between the ages of 50 and 85 (Freedman et al., 2009a), and the Radiation Risk Assessment Tool estimates an individual's lifetime risk of cancer from exposure to ionizing radiation (de Gonzalez et al., 2012). These examples only represent a small number of the many risk calculators available online for providing doctors and patients with more personalized estimates of disease risk. The large number of absolute risk models now directly in use by the public speaks to a growing trend toward managing one's health behaviors in the context of quantitative estimates of disease risk.

In this thesis, we address several methodological issues in developing absolute risk models, dealing with complexities for combining information from various different

sources of data. While the methods are developed with the aim of building models for predicting risk of breast cancer, the methodological issues addressed are applicable in a wide variety of settings. In the following, we describe the two main motivating problems and present the challenges involved with each. First, we discuss the objective of building a subtype-specific absolute risk model for breast cancer by integrating different data sources and we present key data applications. Second, we introduce the problem of how to best make use of external information or existing models in order to calibrate a new risk model more generally.

1.1 Developing Subtype-Specific Absolute Risk Models for Breast Cancer

Subtype-specific models are particularly relevant for breast cancer as it is a heterogeneous disease, encompassing numerous subtypes based on tumor characteristics such as the presence of hormone receptors or growth factors (Burstein, 2005). Distinct breast cancer subtypes differ with respect to age of diagnosis, risk factors, prevention options, treatment regimes, and survival outcomes (Anderson and Matsuno, 2006; Visvanathan et al., 2009; Putti et al., 2005; Burstein, 2005). Researchers who note this heterogeneity recommend taking a stratified approach and express hope that subtype-specific risk prediction may result in better implementation of prevention strategies and earlier tumor detection (Anderson and Matsuno, 2006; Yang et al., 2011a). The methods we develop in this thesis are driven by consideration for the characteristics of real data sources available for building such a model. Specifically, we consider the utility of existing cohort studies, case-control studies, national disease registries, and national surveys for fitting a subtype-specific absolute risk model for breast cancer in the US population.

Cohorts

To fit a subtype-specific absolute risk model, one needs information on covariate relative risks, age-specific disease incidence rates, and covariate and subtype distributions that are representative of the population of interest. In an ideal world, one might have access to an existing large, representative, prospectively-collected cohort study which could provide information for each component of the model. However, cohort studies are expensive to conduct and, especially for rare outcomes such as breast cancer subtypes, require many years of follow-up in order to accrue a large number of cases. In practice, the ideal cohort study for a given application may be unavailable and conducting one may not be feasible or timely. Additionally, a given cohort study may not be representative of incidence in the population as increased intensity of screening and follow-up in the study may artificially produce higher rates than would naturally occur in the population. To mitigate these limitations, we propose methods that enable the use of existing cohort data in conjunction with other complementary data sources, such as registries that contribute more representative incidence rates.

Case-Control Studies

For developing a subtype-specific absolute risk model, case-control studies have some advantages over cohort studies, along with some limitations. Case-control studies specifically recruit cases with the disease of interest so they require less time to obtain a large number of cases, particularly for rare subtypes, and are generally less expensive to conduct than cohort studies. Incident case-control studies can provide representative estimates of the disease subtype distribution, provided the sampled cases are representative of all cases in the population for whom we intend to estimate subtype-specific absolute risk. Case-control studies can be used to estimate covariate

odds ratios, which approximate relative risks for rare diseases such as breast cancer subtypes.

A drawback of case-control studies is that data on lifestyle factors may be affected by selection or recall bias, and one must carefully consider whether a given sample is representative of the underlying population. In particular, a major challenge in designing case-control studies is ensuring that the sampled controls are representative of unaffected individuals in the population. It is well known that hospital-based incident case-control studies are likely to have non-representative controls that are generally less healthy than the overall population; however, such a study may be an excellent source of cases with subtype information. We address the issue of how one can utilize these studies, for which no adequate controls are available, in order to estimate key components of a subtype-specific absolute risk model.

A further limitation of case-control studies is that they do not provide estimates of disease incidence due to the fact that specific numbers of cases and controls are purposefully sought for inclusion in the study. For this reason, one cannot build an absolute risk model from a case-control study alone, even one with good controls. Thus, in order to make use of the rich data provided by case-control studies, we develop methods for combining information with other data sources that can provide the representative incidence rates and, in some cases, the representative controls that case-control studies lack.

Registries

National disease registries are an excellent source of incidence rate information for absolute risk models. A strength of this data is that it is representative of the national population and typically includes immense sample sizes, which result in very precise estimates of incidence. However, these large databases generally collect very

little covariate information, so the reported incidence rates are based on a mixture of women with different covariate levels. For diseases with established subtypes, national registries typically collect the necessary disease characteristic information for delineating subtype-specific incidence rates; however, this is not always the case.

We develop methods for calibrating subtype-specific absolute risk models to nationally representative registry incidence rates, dealing with the situation where the registry is missing some or all tumor characteristics needed to define the subtypes of interest. For many diseases, the research community is still learning how to identify clinically relevent subtypes. In the event that new biomarkers are identified in future research, these methods will allow researchers to calibrate absolute risk models for the newly identified subtypes to incidence rates from a registry that has not yet begun to collect information on the new biomarkers.

Other Challenges

In addition to addressing considerations of study design and missingness, a major statistical component of this research is the development of methods for variance estimation that account for the integration of different data sources. Generally, we approach this problem by applying the functional delta method and concepts from empirical process theory to derive variance estimators. After working through the statistical theory, we computationally implement the proposed methods, validate their performance using simulations, and finally apply them to a real and relevant data example, that of building a subtype-specific absolute risk model for breast cancer.

1.2 Applications

In this thesis, we present two major data applications; specifically, we develop absolute risk models for breast cancer in the US population using data from two large constortia. First, we build an absolute risk model for overall breast cancer using prospective cohort data from the Breast and Prostate Cancer Cohort Consortium (BPC3) (Hunter et al., 2005; Husing et al., 2012). In a second data application, we apply our methods to case-control studies from the Breast Cancer Association Consortium (BCAC) in order to build subtype-specific absolute risk models for breast cancer subtypes defined by estrogen receptor status (Breast Cancer Association Consortium, 2006; Yang et al., 2011b). In both applications we aim to develop models that are representative of the US population. To better accomplish this, we calibrate the models to nationally representative breast cancer incidence rates from the National Cancer Institute's Surveillance Epidemiology and End Results (SEER) database. Much of our subtype-specific methods development is motivated directly by practical considerations encountered in working with these datasets. In the following, we give a brief overview of these three influential data sources.

BPC3

The BPC3 includes 8 large, prospective cohorts that together total more than 17,000 cases and 19,000 controls with breast cancer outcomes (Institute, 2014). Specifically, the consortium includes the following studies conducted in the US population: the American Cancer Society Cancer Prevention Study-II (CPS-II); Harvard's Nurses' Health Study (NHS) and Women's Health Study (WHS); the National Cancer Institute's Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial; the Multiethnic Cohort (MEC); and the Women's Health Initiative (WHI) (Milne

et al., 2010; Fasching et al., 2012; Nickels et al., 2013). It also includes cohorts from Europe and Australia, the European Prospective Investigation of Cancer (EPIC) and the Melbourne Collaborative Cohort Study (MCCS) respectively. The BPC3 data includes well known anthropomorphic and reproductive risk factors for breast cancer, as well as genetic information on 24 single nucleotide polymorphisms (SNPs), which have been previously shown to be associated with breast cancer.

BCAC

The BCAC includes more than 60 case-control studies from many different countries, with a total of more than 90,000 breast cancer cases and 90,000 controls. The BCAC data also includes information on tumor characteristics, including estrogen receptor (ER) status, progesterone receptor (PR) status, and human epidermal growth factor receptor 2 (HER2), biomarkers which are frequently used to classify breast cancer into subtypes. Additionally, the BCAC contains data on standard risk factors for breast cancer along with genetic information on 77 SNPs previously shown to be associated with breast cancer. The case-control studies that make up this consortium include those that are population-based, hospital-based, and family-based, and some studies that are of mixed design. With so many studies, there is a huge variety in the magnitude and patterns of missing data, in both the covariates and the tumor characteristics. In our methodological work, we carefully consider the ramifications of these practical data issues in the context of building subtype-specific absolute risk models and discuss statistical approaches to handle those issues.

SEER

The SEER database monitors cancer incidence in regions across the country covering approximately 28 percent of US population, roughly 88 million people (Howlader et al., 2011). For regions covered by the registry, SEER data includes breast cancer incidence and population rates linked with age, race, sex, registry location, and year of diagnosis (Howlader et al., 2011). Beginning in 1990, SEER began to collect ER and PR status on breast cancer cases, and in 2009 began collecting HER2 information (Fritz and Ries, 1998; Adamo et al., 2011). To be specific, in our data applications we use age-specific incidence rates for integer ages 0 to 85, for those classified as white race, from the SEER 18 Research Data, which contains the largest geographic coverage available in SEER. The age-specific incidence rates were stratified into three rates according to estrogen receptor status as ER positive, ER negative, and ER unknown. In applications where only the overall rates of breast cancer were needed, we simply calculated the age-specific incidence rates for any breast cancer as the sum of these three rates for each age.

1.3 Calibrating to External Information

As new risk factors are identified, there is a need to update existing risk models to fully use the most up-to-date information in predicting disease risk. Ideally, such a risk model should be developed from a large prospective cohort study that is representative of the underlying population and has information on all risk factors, including both the existing and the new ones. In practice, however, such studies are hard to come by. For example, many risk models for cancers and heart diseases have been built using long-established cohort studies; however, these studies are unlikely to have collected the information necessary for evaluating new risk factors which have only recently

been identified.

Rather than conducting an entirely new cohort study to add a few new risk factors to a given risk model, it is more efficient to incorporate information from existing models as much as possible. For many diseases, established risk models have already been developed based on data from large representative cohorts and thoroughly validated in independent studies. Calibrating to this sort of quality information can result in an updated model that is more representative of the underlying population. An example of this, which we have discussed previously, is the idea of calibrating an absolute risk model to national registry data in order to make the model more nationally representative.

Often, case-control studies are the primary source of information on the new risk factors, such as genetic markers, with which to update a model. However, we know that in case-control studies lifestyle factors may be affected by potential selection or recall bias, making them less than ideal for estimating those effects. In this situation, one could benefit by taking advantage of an existing published model, which does have representative information for the existing risk factors that may not be well measured in a given case-control study.

Even if representative information is available on all risk factors from a single new study, building the updated model solely based on that study may be inefficient in that doing so makes no use of the information from published models. If established models were originally built based on a larger, more representative study and carefully evaluated through independent validation studies, one ought to leverage information from those models to develop an updated risk model that is better calibrated to the underlying population.

In this thesis, a key problem we seek to address is exactly how to best use this

external data when building a risk model. We explore the use of a "regression calibration" approach that is popular in the sample-survey community as a tool for increasing efficiency of parameter estimation from a given survey by utilizing information from large external data sources on related variables (Wu and Sitter, 2001; Wu, 2003). We investigate the performance of the estimator both theoretically and numerically, in a setting to which it has not previously been applied, where the external information is representative of the population of interest while the sample population may differ in some respects. We identify conditions under which the calibration estimator is an effective method for calibrating a new risk model to external information, resulting in a model that is calibrated to the external population of interest. We also diagnose some situations where the method does not perform well, and make practical recommendations regarding when the method can be safely applied. Finally, we identify areas of future research, inspired by the goal of improving upon the weaknesses of the proposed calibration estimator.

Chapter 2

Methods for Developing Absolute Risk Models for Disease Subtypes: Integrating Multiple Data Sources

In this chapter we develop a general framework for modeling the absolute risk of disease subtypes. We present the model and discuss a number of different ways to estimate key parameters depending on which data sources are available. We review the strengths of the different approaches and extend the methods to handle situations where there is missing data in the tumor characteristics which define the disease subtypes of interest. Finally, we use empirical process theory to derive variance estimation procedures that account for multiple sources of uncertainty.

2.1 The Model

Absolute risk is the probability that a specific event will occur within a defined interval of time, accounting for the risk of competing events. In this thesis we model absolute

risk where the event of interest is tumor diagnosis. Specifically, we consider the setting where individuals are at risk of being diagnosed with tumor subtypes $j = 1, ..., J$ defined by a set of tumor characteristics, $h = 1, ..., H$. Mathematically speaking, we model the probability that an individual with covariates Z, who is tumor free at age a, is diagnosed with a tumor of subtype j^*, within some upcoming time interval τ, given that the individual is at risk of competing events including other tumor subtypes, $P[a \leq T < a + \tau, j = j^* | T \geq a, Z]$. In the following, we extend the absolute risk model presented in Gail (2011) to accommodate disease subtypes, building up from basic probability principles to provide a clear depiction of the framework upon which the model is based.

The probability of an event in the interval $(a, a + \tau)$ is simply the probability of the event happening at any given time, integrated over the entire interval. Thus we express

$$\begin{aligned} P[a \leq T < a + \tau, j = j^* | T \geq a, Z] &= \int_a^{a+\tau} P[T = t, j = j^* | T \geq a, Z] dt \\ &= \int_a^{a+\tau} P[T = t, j = j^* | T \geq t, Z] \frac{P[T \geq t | Z]}{P[T \geq a | Z]} dt. \end{aligned}$$

We have decomposed the absolute risk probability into an integral over the product of two standard functions in survival analysis: the survival function, $S(t) = P(T \geq t)$, and the subtype specific hazard function, $\lambda_{j^*}(t) = P(T = t, j = j^* | T \geq t)$ (Kalbfleisch and Prentice, 1980). We leverage the classic relationships between these quantities (namely $S(t) = exp(-\Lambda(t))$ and $\Lambda(t) = \int_0^t \lambda(t)$) to further manipulate the expression

of the absolute risk

$$P[a \leq T < a+\tau, j = j^* | T \geq a, Z] = \int_a^{a+\tau} \lambda_{j^*}(t|Z) \frac{S(t|Z)}{S(a|Z)} dt$$

$$= \int_a^{a+\tau} \lambda_{j^*}(t|Z) exp\left(\Lambda(a|Z) - \Lambda(t|Z)\right) dt$$

$$= \int_a^{a+\tau} \lambda_{j^*}(t|Z) exp\left(-\int_a^t \lambda(u|Z) du\right) dt.$$

$\lambda(t|Z)$ is the overall hazard function for any event, including diagnosis of any tumor subtype, as well as competing events. Thus, if we denote the hazard for competing mortality events by $c(t|Z)$, the overall hazard $\lambda(t|Z) = \left(\sum_{j=1}^{J} \lambda_j(t|Z)\right) + c(t|Z)$. This relationship allows us to express the absolute risk solely in terms of the subtype specific hazard functions λ_j and the hazard of competing mortality

$$P[a \leq T < a+\tau, j = j^* | T \geq a, Z] =$$

$$\int_a^{a+\tau} \lambda_{j^*}(t|Z) exp\left(-\int_a^t \left[\sum_{j=1}^{J} \lambda_j(u|Z) + c(u|Z) du\right]\right) dt.$$

Thus far we have not made any modeling assumptions; we have simply worked with standard relationships in probability to express the absolute risk of subtype j^* in terms of subtype specific hazard functions. At this point, we incorporate a Cox proportional hazards model for the subtype specific hazard functions, $\lambda_j(t|Z) = \lambda_{0j}(t) e^{Z\beta_j}$ for $j = 1, ..., J$ where $\lambda_{0j}(t)$ is the baseline hazard function, or the hazard function for an individual with referent level covariates. We also assume that competing mortality

risks do not depend on covariates other than age. These modeling choices result in

$$P[a \leq T < a+\tau, j = j^* | T \geq a, Z] = \tag{2.1}$$

$$\int_a^{a+\tau} \lambda_{0j^*}(t) e^{Z\beta_{j^*}} exp\left(- \int_a^t \left[\sum_{j=1}^{J} \lambda_{0j}(u) e^{Z\beta_j} + c(u) du \right] \right) dt.$$

This is our primary subtype specific absolute risk model of interest. Later we will discuss situations where one might choose to incorporate additional modeling assumptions, but in each case we will use this model as a starting point.

2.1.1 Model Parameters

Fitting this subtype-specific absolute risk model requires estimation of:

1. the hazard rate of competing mortality events $c(t)$
2. the hazard ratio parameter for each subtype β_1, ..., β_J, and
3. the baseline hazard function for each subtype $\lambda_{01}(t)$, ..., $\lambda_{0J}(t)$.

We allow these functions to be as flexible as possible by parametrizing the baseline hazard functions and the competing hazard function non-parametrically, assigning a parameter to each time point where we have data with which to estimate the value of the function. We expand on the details in Section 2.3.

Depending on the available data sources, one can estimate these quantities in a number of different ways. For instance, estimates of the hazard rate of competing mortality events $c(t)$ can be obtained empirically from a representative cohort study with mortality data or given by the age-specific mortality rates provided by a national survey, such as the US National Vital Statistics System. In the following sections, we

discuss a variety of methods for fitting the final two key pieces of the model, starting with the Cox hazard ratio parameters for each subtype.

2.2 Estimating Subtype-Specific Hazard Ratios

In this section, we discuss methods for estimating the subtype-specific Cox hazard ratio parameters from cohort and case-control studies. The exponentiated parameters, e^{β_j}, are subtype-specific hazard ratios. We present options for estimating the β_j parameters in the case where there is missing data in the tumor characteristics variables that define each subtype.

2.2.1 Estimating Hazard Ratios from Cohort Data

Suppose we have data from a cohort study that prospectively follows individuals until they die, are censored, or are diagnosed with a tumor. Specifically, suppose data is collected on (Z, T, S) where Z denotes multivariate covariate data, T the time to the observed event and S the type of event, with $S = 1, ..., J$ for the tumor subtypes and $S = 0$ for death or censoring.

Methods are well established for estimating Cox model hazard ratio parameters from time to event data collected by cohort studies (Kalbfleisch and Prentice, 1980). However, in the absolute risk setting, the different tumor subtypes act as competing risks for one another. The standard method for estimating a particular β_{j^*} in the presence of competing risks was developed by Holt and expanded on by Prentice in 1978. The method is based on a partial likelihood that conditions on subtype. Specifically, the partial likelihood is constructed from the probability that individual i fails at time $t_{(i)}$, conditioning on the risk set at that time ($R_{t_{(i)}} = l : t_l \geq t_{(i)}$) and the fact that exactly one failure *of type* j^* occurs at that time. The partial likelihood

is the product of this probability for all individuals (i) who experienced the event of interest (in this case subtype j^*),

$$PL\left(\beta_{j^*}\right) = \prod_{(i)} \frac{\lambda_{j^*}(t_{(i)}|z_i)}{\sum_{l \in R_{t_{(i)}}} \lambda_{j^*}(t_{(i)}|z_l)} = \prod_{(i)} \frac{\lambda_{0j^*}(t_{(i)}) e^{\beta_{j^*} z_i}}{\sum_{l \in R_{t_{(i)}}} \lambda_{0j^*}(t_{(i)}) e^{\beta_{j^*} z_l}}$$

$$PL\left(\beta_{j^*}\right) = \prod_{(i)} \frac{e^{\beta_{j^*} z_i}}{\sum_{l \in R_{t_{(i)}}} e^{\beta_{j^*} z_l}} \quad \text{(Holt, 1978; Prentice et al., 1978).} \tag{2.2}$$

Conditioning on subtype is useful because the baseline hazard functions cancel out, resulting in a partial likelihood and corresponding score function that only involve the parameter of interest, β_j^*. This partial likelihood results in the score function

$$S\left(\beta_{j^*}\right) = \frac{\partial log PL\left(\beta_{j^*}\right)}{\partial \beta_j} = \sum_{(i):S_i = j^*} \left[z_i - \frac{\sum_{l \in R_{t_{(i)}}} z_l e^{\hat{\beta}_{j^*} z_l}}{\sum_{l \in R_{t_{(i)}}} e^{\hat{\beta}_{j^*} z_l}} \right] = 0. \tag{2.3}$$

Iterative methods are used to obtain the estimate $\hat{\beta}_j$ that solves this equation.

Alternatively, one can estimate β_j^* from a partial likelihood that does not condition on subtype. This unconditional partial likelihood is based on the probability that individual i fails at $t_{(i)}$ given the risk set at time $t_{(i)}$ and that exactly one failure occurs at that time, *regardless of the subtype* (Kalbfleisch and Prentice, 1980). For each time of an observed event $t_{(i)}$, this probability can be expressed in terms of the overall hazard function as the partial likelihood

$$PL'\left(\beta_1, ..., \beta_J\right) = \prod_{(i)} \frac{\lambda(t_{(i)}|z_i)}{\sum_{l \in R_{t_{(i)}}} \lambda(t_{(i)}|z_l)}.$$

The overall hazard of failure $\lambda(t_{(i)}|z_i)$ is comprised of the sum of subtype-specific hazards, where "subtype" $S = 0$ captures the competing non-tumor events. Incorporating the subtype-specific proportional hazards model results in the following expression for the partial likelihood,

$$PL'(\beta_1, ..., \beta_J) = \prod_{(i)} \frac{\lambda(t_{(i)}|z_i)}{\sum_{l \in R_{t_{(i)}}} \lambda(t_{(i)}|z_l)} = \prod_{(i)} \frac{\sum_{j=1}^{J} \lambda_c(t_{(i)}|z_i)}{\sum_{l \in R_{t_{(i)}}} \sum_{j=1}^{J} \lambda_c(t_{(i)}|z_l)} = \prod_{(i)} \frac{\sum_{j=1}^{J} \lambda_{0j}(t_{(i)}) e^{\beta_j z_i}}{\sum_{l \in R_{t_{(i)}}} \sum_{j=1}^{J} \lambda_{0j}(t_{(i)}) e^{\beta_j z_l}}.$$

The estimate $\tilde{\beta_{j^*}}$ from this unconditional partial likelihood solves the score function

$$S(\beta_{j^*}) = \sum_{(i)} \frac{\lambda_{0j^*}(t_{(i)}) z_i e^{\beta_{j^*} z_i}}{\sum_{j=1}^{J} \lambda_{0j}(t_{(i)}) e^{\beta_j z_i}} - \frac{\sum_{l \in R_{t_{(i)}}} \lambda_{0j^*}(t_{(i)}) z_l e^{\beta_{j^*} z_l}}{\sum_{l \in R_{t_{(i)}}} \sum_{j=1}^{J} \lambda_{0j}(t_{(i)}) e^{\beta_j z_l}} = 0. \tag{2.4}$$

Unlike in the score function for the conditional partial likelihood given by equation (2.3), the score function given by equation (2.4) involves the β_j parameters and the baseline hazard functions $\lambda_{0j}(t)$ for all subtypes. This makes parameter estimation slightly more computational in that we must successively update each parameter, plugging in current best versions of the other parameters (including the many parameters defining the non-parametric baseline hazard function), in order to iteratively obtain final estimates. However, we show in Appendix A that if we plug-in a non-parametric estimate of hazard into the unconditional partial likelihood score equation and assume no tied failure times, this equation simplifies to the score equation from the conditional partial likelihood. Thus, under a non-parametric model for baseline hazard, $\hat{\beta}_j$ is equivalent to $\tilde{\beta}_j$. If we were to incorporate a parametric model for baseline hazard, we would expect increased efficiency in the estimate of $\tilde{\beta}_j$ due to the

fact that the unconditional partial likelihood approach, $PL'(\beta_1, ..., \beta_J)$, does not lose information by conditioning.

2.2.2 Estimating Hazard Ratios from Case-Control Data

Appropriate cohort data is not always available and when this is the case another option is to estimate the subtype-specific hazard ratio parameters β_j from case-control data. For rare outcomes, case-control studies are typically more economical in providing a greater number of cases, especially when particular subtypes are of interest. Prentice and Breslow (1978) show how to estimate hazard ratio parameters β from a Cox proportional hazards model using case-control data, and they extend discussion of the method to the competing risk context. The authors demonstrate that the conditional likelihood for β from a retrospective sampling scheme, conditioning on the numbers of cases and controls selected, is the same as the likelihood for prospective data associated with the Cox model (Prentice and Breslow, 1978). Having established this link, they describe a computational strategy for estimating β through standard logistic regression that has the form

$$log\left\{\frac{P[S=j|T=t,z]}{P[S=0|T=t,z]}\right\} = \alpha(t) + z\beta.$$

Effectively, the authors show that one can estimate Cox hazard ratio parameters β by simply fitting a logistic regression model that includes a non-parametric function of time, $\alpha(t)$. In practice, this can be accomplished by including categorical age strata in the model. Logistic regression is standard in statistical software packages, so implementing this method is straightforward. We can apply this method to obtain subtype-specific hazard ratio parameters β_j from case-control data simply by fitting separate logistic regression models for each of the J subtypes.

However, rather than modeling each of the subtype outcomes individually, we propose extending the ideas of Prentice and Breslow (1978) to multinomial logistic regression and instead model the risk of each subtype simultaneously. This is a standard way to model categorical outcomes and it is well known that multinomial logistic regression produces more efficient parameter estimates than individual logistic regression models that estimate the same parameters (Agresti, 2002).

Additionally, because multinomial logistic regression estimates all parameters simultaneously, in certain situations it is possible to fit a multinomial model using data that would necessarily be excluded when fitting a logistic regression model. In section 2.2.3, we describe instances where this is the case and demonstrate how to reparametrize the multinomial model in order to make use of this data.

2.2.3 The Multinomial Likelihood

Several data features of the Breast Cancer Association Consortium (BCAC) motivated us to develop and apply a reparametrized multinomial logistic regression model in order to estimate the hazard ratios for subtypes. Our first consideration focused on how to handle data from the numerous hospital-based studies in the consortium. While the hospital-based case-control studies provided a representative sample of cases from the population, the hospital-based controls did not constitute a representative sample of non-cases in the population. Thus, for the purposes of the analysis, the hospital-based case-control studies could only contribute the representative case data, with the non-representative controls excluded. Additionally, due to the fact that the BCAC data is made up of multiple studies, we needed to estimate covariate hazard ratios adjusted for study to help ensure that the observed relationships were not driven by systematic differences in the way data was collected or processed at different study centers.

However, for the hospital-based studies contributing cases but not controls, it is not possible to fit a model that adjusts for study using standard methods. To estimate the study effect, the estimation procedures in logistic regression and multinomial logistic regression rely on contrasts between cases and controls within a given study, so when a study does not have controls, the study effect is not estimable. Excluding the hospital-based studies entirely would allow use of standard methods, but at the cost of reduced sample size, less efficient parameter estimation, and a colossal waste of data.

Hospital-based studies with cases of different subtypes have information on case-case hazard ratios. If we were to fit logistic regression models for each subtype separately, there would be no way to incorporate that information. However, because multinomial logistic regression estimates parameters for all subtypes simultaneously, in principle it should be possible for hospital-based cases to contribute case-case information to the estimation, with population-based studies that have representative controls contributing to estimation of the case-control parameters. To achieve this we needed to reparametrize the multinomial model.

In the following section, we show how to reparametrize the multinomial logistic regression model in order to include data from studies that are missing controls while appropriately adjusting for study effects. We go on to extend the method to allow cases with incomplete subtype information to contribute to estimation of the model parameters. While the development of this model is motivated by issues encountered in the BCAC data, the reparametrized model should be useful for addressing similar issues that may arise in other datasets as well.

Notation

For data (Z_l, S_l) on individuals $l = 1, ..., N$, let Z_l denote a vector of covariates and let S_l denote an integer disease status, taking values $S_l = 0$ for controls and $S_l = 1, ..., J$ mutually exclusive disease subtypes.

A standard multinomial logistic model for this data takes the form

$$log(\frac{P(S = j|Z)}{P(S = 0|Z)}) = Z\beta_j \tag{2.5}$$

for $j = 1, ..., J$, where β_j is a parameter vector. Standard methods obtain parameter estimates by maximizing the likelihood

$$L = \prod_{i=1}^{N} \prod_{j=0}^{J} P(S_l = j|Z_l)^{I_{\{S_l=j\}}} \tag{2.6}$$

$$L = \prod_{i=1}^{N} \frac{\prod_{j=1}^{J} exp(Z_l\beta_j)^{I_{\{S_l=j\}}}}{1 + \sum_{j=1}^{J} exp(Z_l\beta_j)}. \tag{2.7}$$

However, this multinomial likelihood is currently defined such that the controls are the referent outcome category. As discussed, under this parametrization it is not possible to estimate study effects for the studies that do not have controls.

Assume for now that the studies without controls have cases of all subtypes $j = 1, ..., J$. Data from these studies can be used to estimate

$$log\left(\frac{P(S = j^*|Z)}{P(S = j'|Z)}\right) = log\left(\frac{P(S = j^*|Z)}{P(S = 0|Z)} \Big/ \frac{P(S = j'|Z)}{P(S = 0|Z)}\right) = Z(\beta_j^* - \beta_j')$$

for any $j^*, j' \in 1, ..., J$. Thus, while studies without controls cannot contribute directly to the estimation of the β_j parameters, data from these studies can contribute

to estimation of the contrasts $\beta_{j^*} - \beta_{j'}$ for $j^*, j' \in 1, ..., J$.

Reparametrizing the Likelihood

To leverage this idea in the likelihood, consider the data into two parts (Z_i, S_i) and (Z_k, S_k), with observations from studies with controls indexed by $i = 1, ..., N_i$ and those from studies without controls indexed by $k = 1, ..., N_k$. We know that there are no controls in (Z_k, S_k), so we construct the likelihood conditional on the fact that $S_k \geq 1$. Mirroring equation (2.6), the likelihood for the two part data is

$$L = \left(\prod_{i=1}^{N_i} \prod_{j=0}^{J} P(S_i = j|Z_i)^{I_{\{S_i=j\}}} \right) \left(\prod_{k=1}^{N_k} \prod_{j=1}^{J} P(S_k = j|Z_k, S_k \geq 1)^{I_{\{S_k=j\}}} \right).$$

Again incorporating the multinomial model given by equation (2.5), we express the likelihood in terms of model parameters as

$$L = \left(\prod_{i=1}^{N_i} \frac{\prod_{j=1}^{J} exp(Z_i\beta_j)^{I_{\{S_i=j\}}}}{1 + \sum_{j=1}^{J} exp(Z_i\beta_j)} \right) \left(\prod_{k=1}^{N_k} \frac{\prod_{j=2}^{J} exp(Z_k(\beta_j - \beta_1))^{I_{\{S_k=j\}}}}{1 + \sum_{j=2}^{J} exp(Z_k(\beta_j - \beta_1))} \right).$$

As previously discussed, in this likelihood the studies without controls contribute to estimation of $\beta_j - \beta_1$. To use this in practice, all that remains is to implement a Newton-Raphson algorithm to optimize the likelihood with respect to β_j. This could be done from the existing likelihood directly, but our preference is to reparametrize

the likelihood by defining $\theta_1 = \beta_1$,$\theta_j = \beta_j - \beta_1$ for $j \geq 2$, resulting in

$$L = \left(\prod_{i=1}^{N} \frac{exp(Z_i\theta_1)^{I_{\{S_i=1\}}} \prod_{j=2}^{J} exp(Z_i(\theta_j+\theta_1))^{I_{\{S_i=j\}}}}{1 + exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i(\theta_j+\theta_1))} \right) \left(\prod_{k=1}^{N_k} \frac{\prod_{j=2}^{J} exp(Z_k\theta_j)^{I_{\{S_k=j\}}}}{1 + \sum_{j=2}^{J} exp(Z_k\theta_j)} \right).$$

The score functions

$$S(\theta_1) = \sum_{i=1}^{N_i} Z_i \left(\left[\sum_{j=1}^{J} I_{\{S_i=j\}} \right] - \frac{exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i\,(\theta_j+\theta_1))}{1 + exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i\,(\theta_j+\theta_1))} \right), \quad (2.8)$$

$$S(\theta_{j^*}) = \sum_{i=1}^{N_i} Z_i \left(I_{\{S_i=j^*\}} - \frac{exp(Z_i\,(\theta_{j^*}+\theta_1))}{1 + exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i\,(\theta_j+\theta_1))} \right) \quad (2.9)$$

$$+ \sum_{k=1}^{N_k} Z_k \left(I_{\{S_k=j^*\}} - \frac{exp(Z_k\,(\theta_{j^*}))}{1 + \sum_{j=2}^{J} exp(Z_k\,(\theta_j))} \right)$$

have the familiar form $Z(S-E[S])$. We obtain the maximum likelihood estimator $\hat{\theta} = \begin{bmatrix} \hat{\theta_1} \\ \hat{\theta_j} \end{bmatrix}$ via the Newton-Raphson method by iterating $\hat{\theta}_{new} = \hat{\theta}_{old} - \left[\, I(\theta) \,\right]^{-1} \begin{bmatrix} S(\theta_1) \\ S(\theta_j) \end{bmatrix}$ where $\left[\, I(\theta) \,\right] = -E \begin{bmatrix} \frac{\partial S(\theta_1)}{\partial \theta_1} & \frac{\partial S(\theta_1)}{\partial \theta_j} \\ \frac{\partial S(\theta_j)}{\partial \theta_1} & \frac{\partial S(\theta_j)}{\partial \theta_j} \end{bmatrix}$. In practice, we approximate this expectation empirically

with averages

$$\frac{\partial S(\theta_1)}{\partial \theta_1} = -\sum_{i=1}^{N_i} Z_i' P_{1i} \left(1 - P_{1i}\right) Z_i \tag{2.10}$$

$$\frac{\partial S(\theta_1)}{\partial \theta_{j^*}} = \frac{\partial S(\theta_{j^*})}{\partial \theta_1} = -\sum_{i=1}^{N_i} Z_i' P_{2j^*i} \left(1 - P_{1i}\right) Z_i \tag{2.11}$$

$$\frac{\partial S(\theta_{j^*})}{\partial \theta_{j^*}} = -\sum_{i=1}^{N_i} Z_i' P_{2j^*i} \left(1 - P_{2j^*i}\right) Z_i + -\sum_{k=1}^{N_k} Z_k' P_{3j^*k} \left(1 - P_{3j^*k}\right) Z_k, \tag{2.12}$$

and

$$P_{1i} = \frac{exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i\left(\theta_j + \theta_1\right))}{1 + exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i\left(\theta_j + \theta_1\right))}$$

$$P_{2j^*i} = \frac{exp(Z_i\left(\theta_{j^*} + \theta_1\right))}{1 + exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i\left(\theta_j + \theta_1\right))}; \;\; P_{3j^*k} = \frac{exp(Z_k\theta_{j^*})}{1 + \sum_{j=2}^{J} exp(Z_k\theta_j)}.$$

Getting Back Estimable Parameters

Converting maximum likelihood estimates $\hat{\theta}$ back into the original parametrization is straightforward, with $\hat{\beta}_1 = \hat{\theta}_1$ and $\hat{\beta}_j = \hat{\theta}_j + \hat{\theta}_1$ for $j = 2, ..., J$. This can be accomplished in one step using matrix notation $\hat{\beta} = \left[\, A \,\right] \begin{bmatrix} \hat{\theta_1} \\ \hat{\theta_j} \end{bmatrix}$, $A = \begin{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ & \ddots & \\ 0 & 0 & 1 \end{bmatrix} & [\,0\,] \\ \begin{bmatrix} 1 & 0 & 0 \\ & \ddots & \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & 0 \\ & \ddots & \\ 0 & 0 & 1 \end{bmatrix} \end{bmatrix}$.

However, the parameter vector β includes both covariate effects and study effects, for subtypes j=1,...,J. As discussed previously, some study effects $\beta_j[s]$ are not estimable for studies without controls (only the contrasts $\theta_j[s]$ are estimable, which is why we reparametrized the likelihood as we did). Though we may be able to estimate $\theta_j[s]$,

that does not change the fact that those elements of $\beta_j[s]$ are not estimable. When we map the estimates $\hat{\theta}$ back to $\hat{\beta}$, we must indicate which elements are not estimable. In order for a given study effect $\beta_j[s*]$ to be estimable, the study of interest and the referent study must both have controls and cases of subtype j. In general, it is not a hindrance that some study effects are not estimable because in most situations they are simply nuisance parameters in a model where our primary goal is to estimate covariate effects. The main purpose of including study in the model is to adjust for the study effect, which is accomplished under the θ_j parametrization before mapping back to β_j. Thus even though study effects cannot be estimated under the β_j parametrization, the final estimable β_j estimates are appropriately adjusted for study.

Incorporating Incomplete Information for Classifying Case Subtype

Another feature of the BCAC data that we wanted to handle in developing the reparametrized multinomial model was missing data in the subtype defining characteristics. For example, we have discussed that breast cancers can be classified into subtypes based on estrogen receptor (ER) status and progesterone receptor (PR) status, which define four breast cancer subtypes: ER-PR-, ER-PR+, PR+ER-, and ER+PR+. In the BCAC data ER status is missing on 19.4% of the cases and PR status is missing on 30.4%. For cases with missing data on one of the tumor characteristics, only partial information is available for assigning subtype outcome and multiple different subtypes are possible based on the known information. For example, if an individual's breast cancer is known to be ER+ but the PR status is missing, then the tumor could be either ER+PR- or ER+PR+, while we know that it is not ER-PR- or ER-PR+. Though this information is incomplete, it would be wasteful to exclude the partial information, especially on such a large number of cases. In this section, we extend the ideas from the previous likelihood to allow individuals with

incomplete subtype information to contribute to estimation of the model parameters, while still accommodating for studies without controls.

In the previous section, we defined the outcome S as a single scalar taking values 0,1,..., J, depending on that individual's specific combination of observed disease characteristics. This implicitly assumed that there was complete information available on the disease defining characteristics in order to assign each individual's tumor to exactly one case subtype. To handle the fact that multiple subtypes are possible when tumor characteristics are missing, we adapt our notation to define the outcome as a binary vector $S_i = [S_0, S_1, ..., S_j]_i$ according to whether the subtype is a possibility given the known disease defining characteristics. When enough information is available to define the subtype exactly, the S_i vector will only assign one S_{ji} to 1.

In the previous section when we assumed complete subtype information, each S_i only needed to reflect a single known outcome, say j^*. Thus, the likelihood given by equation (2.7) only needed to reflect $P(S_i = j^*|Z_i)$ for each person, given generally as $\prod_{j=0}^{J} P(S_i = j|Z_i)^{I_{\{S_i=j\}}}$. However, when there is incomplete information to determine subtype, the likelihood must account for some individuals whose incomplete tumor information allows multiple tumor subtypes to be possible, say $S_i = j^*$ or j'. In this case, the likelihood should reflect

$$P(S_i = j^* \text{ or } S_i = j'|Z_i) = P(S_i = j^*|Z_i) + P(S_i = j'|Z_i),$$

expressed generally as $\sum_{j=1}^{J} I_{\{S_{ji}=1\}} P(S_i = j|Z_i)$. This results in the likelihood

$$L = \left(\prod_{i=1}^{N_i} \sum_{j=1}^{J} I_{\{S_{ji}=1\}} P(S_i = j|Z_i) \right) \left(\prod_{k=1}^{N_k} \sum_{j=1}^{J} I_{\left\{S_{jk}=1\right\}} P(S_k = j|Z_i, S_k \geq 1) \right).$$

Again we parametrize the likelihood with parameters β_j for $j = 1, ..., J$ from the

multinomial model (2.5), resulting in

$$L = \left(\prod_{i=1}^{N_i} \frac{I_{\{S_{0i}=1\}} + I_{\{S_{1i}=1\}} exp(Z_i\beta_1) + \sum_{j=2}^{J} I_{\{S_{ji}=1\}} exp(Z_i\beta_j)}{1 + \sum_{j=1}^{J} exp(Z_i\beta_j)} \right) \cdot$$

$$\left(\prod_{k=1}^{N_k} \frac{I_{\{S_{1k}=1\}} + \sum_{j=2}^{J} I_{\left\{S_{jk}=1\right\}} exp(Z_k(\beta_j - \beta_1))}{1 + \sum_{j=2}^{J} exp(Z_k(\beta_j - \beta_1))} \right).$$

Again, we reparametrize according to $\theta_1 = \beta_1$ and $\theta_j = \beta_j - \beta_1$ for $j \geq 2$, resulting in

$$L = \left(\prod_{i=1}^{N_i} \frac{I_{\{S_{0i}=1\}} + I_{\{S_{1i}=1\}} exp(Z_i\theta_1) + \sum_{j=2}^{J} I_{\{S_{ji}=1\}} exp(Z_i(\theta_j + \theta_1))}{1 + \sum_{j=1}^{J} exp(Z_i(\theta_j + \theta_1))} \right) \cdot$$

$$\left(\prod_{k=1}^{N_k} \frac{I_{\{S_{1k}=1\}} + \sum_{j=2}^{J} I_{\left\{S_{jk}=1\right\}} exp(Z_k\theta_j)}{1 + \sum_{j=2}^{J} exp(Z_k\theta_j)} \right)$$

As before, the score functions have the form $Z(S - E[S])$:

$$S(\theta_1) = -\sum_{i=1}^{N_i} Z_i \left(I_{\{S_{0i}=1\}} - \frac{1}{1 + exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i(\theta_j + \theta_1))} \right), \tag{2.13}$$

$$S(\theta_{j^*}) = -\sum_{i=1}^{N_i} Z_i \left(I_{\left\{S_{j^*i}=1\right\}} - \frac{exp(Z_i(\theta_{j^*} + \theta_1))}{1 + exp(Z_i\theta_1) + \sum_{j=2}^{J} exp(Z_i(\theta_j + \theta_1))} \right) \tag{2.14}$$

$$-\sum_{k=1}^{N_k} Z_k \left(I_{\left\{S_{j^*k}=1\right\}} - \frac{exp(Z_k\theta_{j^*})}{1 + \sum_{j=2}^{J} exp(Z_k\theta_j)} \right).$$

These score functions are essentially the same as the score functions for complete subtype data given by equations (2.8) and (2.9); however, in this formulation an individual may have more than one non-zero indicator in the outcome vector $S_i = [S_0, S_1, ..., S_j]_i$ if multiple subtypes are possible due to some unknown subtype defining information. The derivatives of these score functions are the same as those given previously in equations (2.10), (2.11), and (2.12), resulting in the same expressions for the Fisher's information matrix and the same Newton-Raphson estimation procedure as previously described.

Calculating the Variance

$\hat{\theta}$ is a maximum likelihood estimator so the covariance $V\left[\hat{\theta}\right] = \left[\, I(\theta) \,\right]^{-1}$, where $I\left(\theta\right)$ is the Fisher's information matrix based on the derivatives of the score functions

(2.13) and (2.14). Given the relationship $\beta = [\, A \,] \begin{bmatrix} \theta_1 \\ \theta_j \end{bmatrix}$, $A = \begin{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ & \ddots & \\ 0 & 0 & 1 \end{bmatrix} & [\,0\,] \\ \begin{bmatrix} 1 & 0 & 0 \\ & \ddots & \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & 0 \\ & \ddots & \\ 0 & 0 & 1 \end{bmatrix} \end{bmatrix}$, the covariance for $\hat{\beta}$ can be easily calculated from the covariance matrix for $\hat{\theta}$ as $V\left[\hat{\beta}\right] = V\left[A\hat{\theta}\right] = A\,[\, I(\theta)\,]^{-1} A'$.

Another way to calculate the variance of θ and β is to use influence functions. This method of variance calculation is based on expressing $\sqrt{N}(\hat{\theta} - \theta) = \frac{1}{\sqrt{N}} \sum_{m=1}^{N} \psi_m$ independent identically distributed contributions and applying the central limit theorem to conclude that the asymptotic variance of $\frac{1}{\sqrt{N}} \sum_{m=1}^{N} \psi_m$, and thus of $\sqrt{N}(\hat{\theta} - \theta)$, is $E_{\theta_0}[\psi' \cdot \psi]$. With the appropriate formula for computing ψ which we derive below, the variance is estimated by the empirical variance of ψ_m in the data.

In this setting, the primary goal is to estimate covariate effects while adjusting for study effects. When we compute the influence functions for the covariate effect parameters, we must take into account the additional variance that arises due to estimation of the nuisance study effect parameters. If we denote the effects of interest θ_p and the nuisance effects θ_{nui}, the function that does this is the efficient influence function, which for maximum likelihood estimators takes the form

$$\psi_{pk} = E_{\theta_p}\left[S(\theta_p)^{eff} S(\theta_p)^{eff\prime}\right]^{-1} S(\theta_p)^{eff}, \quad \text{with}$$
$$S(\theta_p)^{eff} = S(\theta_p) - E_{\theta}\left[S(\theta_{nui})'\right] E_{\theta_{nui}}\left[S(\theta_{nui}) S(\theta_{nui})'\right]^{-1} S_{\theta_{nui}}.$$

We compute the efficient influence functions empirically using the expressions for the

score functions given by equations (2.13) and (2.14) as

$$\psi_{pk} = \left(\frac{1}{N} \sum_{m=1}^{N} S(\theta_p)_m^{eff} S(\theta_p)_m^{eff'} \right)^{-1} S(\theta_p)_m^{eff}$$

$$S(\theta_p)_m^{eff} = S(\theta_p)_m -$$

$$\left(\frac{1}{N} \sum_{m=1}^{N} S(\theta_p)_m S(\theta_{nui})'_m \right) \left(\frac{1}{N} \sum_{m=1}^{n} S(\theta_{nui})_m S(\theta_{nui})'_m \right)^{-1} S(\theta_{nui})_m.$$

Based on the linear relationship between θ and β, once we have the influence functions $\psi_{p(NxP)}$ for the parameters of interest θ_p we can easily obtain the influence function for β_p as $\phi_p = \psi_p A'$. The influence functions ϕ_p are useful for other variance calculations, such as computing the variance of a complex function $f(\hat{\beta}_p)$. Specifically, ϕ_p will later be used in a variance calculation for subtype-specific absolute risk estimates that are based on β_j parameters from this reparametrized multinomial model.

Simulations

In order to eventually incorporate this method into our absolute risk model, we implemented the reparametrized multinomial likelihood in R. To verify that the coded method was functioning well, we evaluated its performance in several simulation settings.

Specifically, we generated cohort data by simulating three predictor variables, distributed as $x_1 \sim N(\mu = 0, \sigma^2 = 9)$, $x_2 \sim Bernoulli(p = 0.2)$, $x_3 \sim Bernoulli(p = 0.5)$, and randomly assigned individuals to one of three studies. We then generated an outcome variable y, taking values 0 for controls or 1, 2, 3, 4 for subtypes, from a multinomial model with known parameters, given in Table 2.1. In this simulation we considered the four subtypes to be defined by two binary tumor characteristics,

Table 2.1: True Parameter Values for Multinomial Simulation: Covariate Log Odds Ratios for Four Subtypes

	Subtype 1	Subtype 2	Subtype 3	Subtype 4
β_{x_0}	-2	-3	-1	-1
β_{x_1}	log(1)	log(1.03)	log(1.04)	log(1.05)
β_{x_2}	log(0.75)	log(1)	log(0.25)	log(0.5)
β_{x_3}	log(0.6)	log(0.6)	log(0.95)	log(0.7)

ER status and PR status. Accordingly, we generated tumor characteristic variables ER and PR based on the simulated subtype variable y, with Subtype 1 = ER-PR-, Subtype 2 = ER-PR+, Subtype 3 = ER+PR-, and Subtype 4 = ER+PR+.

In the first simulation setting, we fit the model with the complete data, with controls in all three studies and complete information on the tumor characteristics. In the second setting, we removed all controls in the second study to evaluate performance in the case where some studies are fully missing controls. In the third setting, we randomly inserted missing values into the ER and PR variables to examine the method's performance when some tumor characteristic information is missing.

In each setting, we generated 500 datasets of sample size 200,000 and used the multinomial likelihood method to estimate the hazard ratio parameters for each subtype. We assigned Subtype 4 (ER+PR+) to be the referent subtype in the reparametrization, both to test our function's option for user specification of the referent subtype and to ensure that the most commonly occuring subtype was used as referent.

Table 2.2 presents the bias as a percentage of the true parameter value for the log odds ratio estimates from the reparametrized multinomial model for under the three simulation settings. For parameters with true value equal to zero, we present the absolute bias instead of the percent bias, which was cannot compute as it would

Table 2.2: Percent Bias of Reparametrized Multinomial Estimates for Covariate Log Odds Ratio Parameters of Four Subtypes, for True Values in Table 2.1

		Percent Bias			
		Subtype 1	Subtype 2	Subtype 3	Subtype 4
Complete Data	β_{x_1}	*2.0E-04	-1.3	-0.2	-0.2
	β_{x_2}	-0.2	*-2.9E-04	0.0	0.0
	β_{x_3}	0.1	-0.6	-0.4	0.1
No Controls in Study 2	β_{x_1}	*-2.6E-05	-0.2	0.1	0.1
	β_{x_2}	-0.1	*-9.9E-04	0.1	0.1
	β_{x_3}	0.0	-0.5	-0.6	0.2
Missing Data in Tumor Characteristics	β_{x_1}	*-7.1E-05	-0.1	-0.2	-0.4
	β_{x_2}	0.4	*3.8E-04	-0.1	-0.1
	β_{x_3}	0.0	0.4	0.7	0.0

" * " indicates where bias is reported instead of percent bias due to true values of 0.

involve dividing by zero. The results in Table 2.2 show that the implemented method provides unbiased parameter estimates in all simulation settings.

Having verified the unbiasedness of the parameter estimates, we next evaluate the performance of the variance estimators. We will refer to the Fisher's information variance estimator as "model-based" and the variance estimator derived from the influence functions as "robust." We compare the standard deviations from these variance estimators to the average empirical standard deviation in the point estimates observed for the 500 datasets. In Table 2.3 we present the bias and coverage probability for the model-based and robust estimates of standard deviation in the three simulation settings . The results in Table 2.3 show that both the model-based and robust variance estimators have less than 10% bias in all cases and coverage probabilities consistently at the 95% level. We see a small percentage of bias in the standard

Table 2.3: Percent Bias and Coverage Probability of Model-Based and Robust Standard Deviation Estimates for Covariate Log Odds Ratio Parameters of Four Subtypes

		Percent Bias				Coverage Probability			
		Y=1	Y=2	Y=3	Y=4	Y=1	Y=2	Y=3	Y=4
Complete Data									
Model-Based	β_{x_1}	-7.3	-5.2	-3.1	-4.7	0.97	0.97	0.96	0.95
	β_{x_2}	4.1	1.7	-1.1	-1.5	0.95	0.95	0.96	0.95
	β_{x_3}	-3.4	0.0	-1.8	1.4	0.97	0.94	0.95	0.94
Robust	β_{x_1}	-7.3	-5.3	-3.1	-4.7	0.97	0.96	0.96	0.96
	β_{x_2}	4.1	1.7	-1.1	-1.6	0.95	0.95	0.96	0.95
	β_{x_3}	-3.4	0.0	-1.9	1.4	0.97	0.94	0.95	0.94
No Controls in Study 2									
Model-Based	β_{x_1}	-6.5	-2.2	2.1	-1.1	0.96	0.97	0.94	0.95
	β_{x_2}	1.7	0.1	-1.1	-1.3	0.94	0.96	0.95	0.95
	β_{x_3}	-1.0	-1.4	-3.5	4.0	0.96	0.95	0.96	0.93
Robust	β_{x_1}	-6.6	-2.3	2.1	-1.1	0.96	0.97	0.94	0.95
	β_{x_2}	1.7	0.1	-1.1	-1.3	0.94	0.96	0.95	0.95
	β_{x_3}	-1.1	-1.4	-3.5	4.0	0.96	0.95	0.96	0.93
Missing Data in Tumor Characteristics									
Model-Based	β_{x_1}	2.6	-2.3	-0.3	-3.5	0.93	0.95	0.95	0.96
	β_{x_2}	-2.7	0.9	-5.7	1.1	0.96	0.94	0.96	0.96
	β_{x_3}	-2.9	3.3	0.1	-4.8	0.95	0.94	0.95	0.96
Robust	β_{x_1}	2.6	-2.3	-0.3	-3.5	0.93	0.96	0.95	0.96
	β_{x_2}	-2.7	0.8	-5.7	1.1	0.96	0.94	0.96	0.96
	β_{x_3}	-2.9	3.2	0.1	-4.8	0.95	0.94	0.95	0.96

"Y" indicates subtype

deviation estimates for the current simulation parameters, with sample size 200,000 and the 'true values' of standard deviation computed empirically on 500 datasets; however, the bias is not systematic. When we compared the model-based standard deviations to the robust ones, we found that they were essentially equal, differing on average by less than 0.005 percent. The model is correctly specified, so this concordance is to be expected and provides further verification that our implementation of the method is working correctly.

We also performed two additional simulations, considering the setting with controls fully missing from one study *and* missing data in the tumor characteristics, as well as a setting with complete data that included an interaction between x_2 and x_3. The results (not shown) did not differ significantly from the three simulations presented here.

2.3 Estimating Baseline Hazard Functions

In addition to the subtype-specific hazard ratio, another key component of a subtype-specific absolute risk model is the baseline hazard function. The subtype-specific baseline hazard function $\lambda_{0j}(t)$ captures the probability $P[T = t, S = j | T \geq t, Z_0]$, over time. At any given time t, this is the chance that an individual is diagnosed with subtype j given that they have not been diagnosed with any subtype prior to that time, for an individual with referent level covariates Z_0. In the following section, we present a number of different options for estimating the subtype-specific baseline hazard function by integrating information from a variety of data sources, including analytic cohort studies, case-control studies, national surveys, and population-based cancer registries. We go on to discuss how one can make use of registry data even if it is lacking detailed subtype and/or covariate information.

2.3.1 Estimating Baseline Hazard Functions from Cohort Data

We begin by reviewing a standard method for estimating subtype-specific baseline hazard functions from cohort data. As in Section 2.2.1, suppose there is a cohort with variables (Z, T, S) where Z denotes multivariate covariate data, T the time to the observed event and S the type of event, with $S = 1, ..., J$ for the tumor subtypes and $S = 0$ for death or censoring. A common way to model the baseline hazard function $\lambda_{0j}(t)$ for each subtype is to do so non-parametrically, allowing the function to be as flexible as possible and driven by the data. Let non-parametric baseline hazard function for each subtype $\lambda_{0j}(t)$ be defined by parameters λ_{jq} at each time point $t_{j_1}, ..., t_{j_{Q_j}}$ where a failure of the given subtype occurred, with zero hazard between these time points, yielding

$$\lambda_{0j}(t) = \begin{Bmatrix} \lambda_{jq} & \text{for } t_{jq} = t \\ 0 & \text{else} \end{Bmatrix} = \sum_{q=1}^{Q_j} \lambda_{jq} I\left\{t_{jq} = t\right\}.$$

This baseline hazard function can be estimated by obtaining maximum likelihood estimates for every parameter λ_{jq} that defines the function.

In Appendix B, we give the details for constructing the likelihood in terms of subtype-specific hazard functions, incorporating the proportional hazards model for each subtype, $\lambda_j(t) = \lambda_{0j}(t)e^{\beta_j z_i}$, and expressing the baseline hazard function $\lambda_{0j}(t)$ in terms of the parameters λ_{jq}, yielding the likelihood

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} \left(\sum_{q=1}^{Q} I\left\{S_i = j\right\} \lambda_{jq} I\left\{t_{jq} = t_i\right\} \right) e^{\beta_j z_i} \right) \cdot \left(\prod_{i=1}^{N} exp\left\{ -\sum_{j=1}^{J} \left(\sum_{m:t_{jm} < t_i} \lambda_{jm} \right) e^{\beta_j z_i} \right\} \right).$$

In Appendix B we also derive the score function for each parameter λ_{jq} and show that the maximum likelihood estimator $\hat{\lambda}_{jq}$ for a specific subtype j^* at time $t_{j^*q^*}$ is

$$\hat{\lambda}_{j^*q^*} = \frac{\sum_{i:S_i=j^*} I\left\{t_{j^*q^*} = t_i\right\}}{\sum_{i:t_{j^*q^*}<t_i}^{N} e^{\beta_{j^*} z_i}} = \frac{d_{j^*}(t_{j^*q^*})}{\sum_{i \in R(t_{j^*q^*})} e^{\beta_{j^*} z_i}}, \tag{2.15}$$

the number of failure events of type j^* which occur at $t_{j^*q^*}$, divided by weighted contributions for all individuals in the risk set $R(t_{j^*q^*})$, those who were event free at $t_{j^*q^*}$. This is equivalent to treating all events other than subtype j^* as censored and applying Breslow's estimator (Breslow, 1972). Consequently, it is straightforward to utilize existing software packages to estimate the baseline hazard function for each subtype by appropriately redefining the censoring variable.

Drawing Strength Across Subtypes

Here, we discuss additional modeling assumptions that could be incorporated to draw strength across all subtypes in the estimation of each subtype-specific baseline hazard functions. We initially defined the subtype-specific Cox proportional hazards models as $\lambda_j(t|Z) = \lambda_{0j}(t)e^{Z\beta_j}$, formalizing the assumption that for a given subtype j the hazard functions for women with different covariate levels are proportional. This does not make any assumptions about the relationship between the baseline hazard functions of different subtypes, say $\lambda_{0j^*}(t)$ and $\lambda_{0j'}(t)$. However, one could choose to model the relationship between the baseline hazard functions of different subtypes, reducing the number of parameters that need to be estimated and leading to gains in the efficiency and stability of the parameter estimates.

Specifically, one could model $\lambda_{0j}(t) = \lambda_{01}(t)h_j(\theta_j, t)$, where $h_j(\theta_j, t)$ is a parametric function and $h_1(\theta_1, t) = 1$. In order to estimate the subtype-specific hazard functions

based on this model, one need only estimate the reference baseline hazard function $\lambda_{01}(t)$ and the parameters θ_j of the functions $h_j(\theta_j, t) = 1$ for j=2,..., j=J. We derive the estimators for this general model.

Again, we characterize the reference baseline hazard function $\lambda_{01}(t)$ by constant parameters $\lambda_1, ..., \lambda_Q$ at all times, $t_1, ..., t_Q$, where an event was observed to occur, with zero hazard in between. Starting with the likelihood from Appendix B, which is defined in terms of the subtype-specific baseline hazard functions, we incorporate the new model $\lambda_j(t) = \lambda_{01}(t)h_j(\theta_j, t)e^{\beta_j z_i}$. We then obtain profile likelihood estimates for $\lambda_{01}(t)$, by treating $h_j(\theta_j, t)$ as fixed. Appendix C contains the mathematical details showing that the estimator at a given time t_{q*} is

$$\hat{\lambda}_{q*} = \frac{\sum_{i:S_i \geq 1} I\{t_{q*} = t_i\}}{\sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{m:t_m < t_i} h_j(\theta_j, t_j)e^{\beta_j z_i}} = \frac{d_{(t_{q*})}}{\sum_{i \in R_{(t_{q*})}} \sum_{j=1}^{J} h_j(\theta_j, t_j)e^{\beta_j z_i}}, \tag{2.16}$$

the total number of observed events that occured at that time, $d_{(t_q)}$, divided by weighted contributions from all those at risk at that time. Recall that the numerator of the baseline hazard estimator for the standard Cox model given in equation (2.15) only included cases of the particular subtype j^*. In contrast, the numerator of the baseline hazard estimator given by equation (2.16) includes all events that occur at t_q, allowing the estimator to draw strength across all subtypes. The estimator also depends on both $h_j(\theta_j, t_j)$ and β_j.

In Appendix D we derive score functions for θ_j parameters from the general model $h_j(\theta_j, t)$. In the case of the simple model, $h_j(\theta_j, t) = \theta_j$, the score function can be

analytically solved to obtain the estimator

$$\hat{\theta}_{j^*} = \frac{d_{j^*}}{\sum_{i=1}^{N} - \int_0^{t_i} \lambda_{01}(u) e^{\beta_{j^*} z_i} du},$$

where d_{j^*} is the number of observed events of subtype j^* irrespective of the time that those events occurred. In both the general formulation of $h_j(\theta_j, t)$ and the simple model where $h_j(\theta_j, t) = \theta_j$, the estimates of $\hat{\theta}_j$ depend on the referent baseline hazard parameters λ_q, so it is necessary to iterate between the estimators in order to obtain final estimates.

2.3.2 Estimating Baseline Hazard Functions from Registry Data

While cohort studies can be an excellent source of representative, prospective data from a population, these studies are expensive and usually conducted over long periods of time. For this reason, when data is needed to fit an absolute risk model for a particular outcome, an ideal cohort study may not be available and often one must resort to case-control studies to estimate the subtype-specific hazard ratio parameters, β_j, using the methods discussed in Sections 2.2.2 and 2.2.3. However, case-control studies cannot be used to estimate the baseline hazard function. In this case, one can use external registry data to estimate baseline hazard functions.

Even if appropriate cohort data is available, there are compelling reasons to consider using registry data, instead of the cohort data, to estimate the baseline hazard functions. Data from a national registry could be more representative of a country's overall population than a single cohort study, especially if the cohort study was conducted in a specialized population. Additionally, the typically large sample size

of registry data can increase efficiency in the estimation of rates, especially for rare cancer subtypes.

Methods for incorporating registry data to estimate the baseline hazard for a single disease outcome have been described by Gail et al. (1989). In particular, the existing methods handle the issue that registries generally collect minimal covariate information and thus report only the marginal hazard rates, $\lambda_m(t) = P[T = t|T \geq t]$. These rates quantify the hazard for individuals with a mixture of different covariate levels, whereas to fit an absolute risk model, an estimate of the hazard function for individuals with referent level covariates $\lambda_0(t) = P[T = t|T \geq t, Z_0]$ is needed. Gail et al. (1989) proposed methods that relate these two quantities through a well-known public health measure, the attributable risk, and used estimates of attributable risk to reweight the marginal hazard rates in order to obtain baseline hazard rates for a single subtype outcome.

In this section, we extend the methods of Gail et al. (1989) to the setting of multiple disease subtypes. First, we describe methods for estimating subtype-specific attributable risks and present the details for naturally extending the reweighting approach described above to handle subtypes. We then address an added complexity in using registry data to estimate baseline hazard rates for disease subtypes: potentially missing information on the subtype-defining tumor characteristics.

To rievew, Bruzzi et al. (1985) define attributable risk as the fraction of the total disease experience in the population that would not have occurred if the effect associated with the risk factor of interest were absent; mathematically,

$$AR(t) = \frac{\lambda_m(t) - \lambda_0(t)}{\lambda_m(t)} = 1 - \frac{\lambda_0(t)}{\lambda_m(t)}.$$

Applying this definition to subtypes, the attributable risk for a given subtype j at

a particular time can be expressed as the difference between the subtype-specific marginal hazard and the subtype-specific hazard for those with baseline covariates as a proportion of the subtype-specific marginal hazard,

$$AR_j(t) = \frac{\lambda_{mj}(t) - \lambda_{0j}(t)}{\lambda_{mj}(t)} = 1 - \frac{\lambda_{0j}(t)}{\lambda_{mj}(t)}.$$

Following Gail et al. (1989), a simple rearrangement of this equation shows that the baseline hazard function, $\lambda_{0j}(t)$, can be obtained from the subtype-specific attributable risk and the marginal hazard function provided by the registry as $\lambda_{0j}(t) = (1 - AR_j(t))\lambda_{mj}(t)$. To make practical use of this relationship, we need to estimate the subtype-specific absolute risks $AR_j(t)$. We will review how to estimate the subtype-specific attributable risk from cases only and from a sample of the population by applying the estimators given by Bruzzi et al. (1985) to subtype outcomes.

Bruzzi's Formula: Estimating Attributable Risk from Cases Only

Bruzzi et al. (1985) showed that as long as we have estimates of the relative risk, we can estimate the attributable risk from the covariate distribution among the cases only. A simple extension of Bruzzi et al.'s estimator to the setting of subtypes yields the estimator

$$\hat{AR}_j = 1 - \frac{1}{d_j} \sum_{i:S_i=j} \frac{1}{\hat{RR}_{ji}} = 1 - \frac{1}{d_j} \sum_{i:S_i=j} \frac{1}{e^{\hat{\beta}_j Z_i}}. \tag{2.17}$$

As before, d_j denotes the total number of subtype j cases. As detailed in Section 2.2, the $\hat{\beta}_j$ estimates can be based on either cohort or case-control data. Based on the estimate in equation (2.17), $\lambda_{0j}(t)$ relates to $\lambda_{mj}(t)$ through a weighting of the marginal hazard function by the expected value of the inverse relative risk among

subtype j cases,

$$\hat{\lambda}_{0j}(t) = \hat{\lambda}_{mj}(t) \left(\frac{1}{d_j} \sum_{i:S_i=j} \frac{1}{e^{\hat{\beta}_j Z_i}} \right) \approx \hat{\lambda}_{mj}(t) \cdot E_{S=j} \left[\frac{1}{e^{\beta_j Z}} \right]. \tag{2.18}$$

The attributable risk link depends only on the distribution of the covariates among subtype j cases, so any data source with a representative sample of cases can be used for estimation. For instance, the covariate distribution among subtype j cases can come from a population-based or hospital-based case-control study, or from a cohort study. However, for any of these studies, it is important that the data constitute a sample of subtype j cases that is representative of the population of interest. If the study is conducted in a special population, this may not be the case and could result in biased estimates.

Estimating Attributable Risk from a Sample of the Population

Again following Bruzzi et al. (1985), another way to obtain the baseline hazard function for each subtype is by weighting the marginal hazard estimates according to the distribution of covariates in the population, rather than just the cases. Starting with the relationship $AR_j(t) = 1 - \frac{\lambda_{0j}(t)}{\lambda_{mj}(t)}$, one can express the marginal hazard for a given subtype as a mixture of subtype-specific hazards which are conditional on covariates and incorporate the proportional hazards model to show that

$$AR_j(t) = 1 - \frac{\lambda_{0j}(t)}{\lambda_{mj}(t)} = 1 - \frac{\lambda_{0j}(t)}{\sum_Z \lambda_j(t|Z) P[Z|T \geq t]} = 1 - \frac{\lambda_{0j}(t)}{\sum_Z \lambda_{0j}(t) e^{\beta_j Z} P[Z|T \geq t]}$$

$$AR_j(t) = 1 - \frac{1}{\sum_Z e^{\beta_j Z} P[Z|T \geq t]}.$$

For rare diseases one may assume that the distribution of covariates for individuals who survive beyond time t is approximately the same as the distribution in the general population, i.e. $P[Z|T \geq t] = P[Z]$, and thus that

$$\tilde{AR}_j(t) = 1 - \frac{1}{\sum_Z e^{\hat{\beta}_j Z} \hat{P}[Z|T \geq t]} \approx 1 - \frac{1}{\sum_Z e^{\hat{\beta}_j Z} \hat{P}[Z]}. \tag{2.19}$$

This places fewer restrictions on the data sources that can be used to estimate the attributable risk. Aside from estimates of the subtype specific hazard ratio parameters, $\hat{\beta}_j$, this calculation depends entirely on estimates of the joint covariate distribution $\hat{P}[Z]$, which can be estimated from any sample with covariate data that is representative of the population of interest, such as a population-based cohort study, a national survey, or the controls from a population-based case-control study. From these data sources, $\hat{AR}_j(t)$ can be obtained empirically or with further modeling of $P[Z]$.

Putting this together with the relationship $\lambda_{0j}(t) = (1 - AR_j(t))\lambda_{mj}(t)$, the subtype-specific marginal hazard functions relate to the baseline hazard functions through the inverse of the expected relative risk in the population,

$$\tilde{\lambda}_{0j}(t) = \frac{\lambda_{mj}(t)}{\sum_Z e^{\hat{\beta}_j Z} \hat{P}[Z]} \approx \frac{\lambda_{mj}(t)}{E_Z\left[e^{\hat{\beta}_j Z}\right]}. \tag{2.20}$$

We have shown how the attributable risk relates the marginal hazard functions provided by the registry to the baseline hazard functions needed in the model. We detailed two methods for estimating the attributable risk, using the distribution of the covariates in a representative sample of cases or of individuals in the population. Both are valid approaches and the decision for which method to use should depend on the quality and representativeness of the available data.

Thus far, we have discussed estimates $\hat{AR}_j$ and $\tilde{AR}_j$ which are constant over time.

A simple but effective way to allow the attributable risk to vary with time is through a piecewise constant function, with each parameter representing the attributable risk within a different stratum of age. The same estimators can be applied, restricting the calculation to covariate data within the appropriate age stratum. In order to do this, the data sources used to estimate time varying attributable risk must have information on age.

Registry Data with Missing Tumor Characteristics

In some cases, a registry may not contain all the information necessary to determine subtype. In the most extreme case, the registry may not contain any information on subtype and thus only provides only the overall hazard rate of any tumor. In this situation, it is still possible to estimate baseline hazard functions for the subtypes of interest calibrated to registry data, by incorporating subtype distribution information from other data sources.

The overall age specific incidence rate provided by the registry, $\lambda_{m+}(t)$, will reflect the hazard for a mixture of subtypes and individuals with different covariate levels. We must obtain $\lambda_{mj}(t)$, the age-specific incidence function for each subtype marginalized over the covariates. Once we have $\lambda_{mj}(t)$, we can apply previously discussed methods that make use of attributable risk to obtain the baseline hazard $\lambda_{0j}(t)$.

One approach to get $\lambda_{mj}(t)$ from $\lambda_{m+}(t)$ would be to reparametrize the marginal hazard of each subtype as a proportional of the overall marginal hazard, $\frac{\lambda_{mj}(t)}{\lambda_{m+}(t)} = \xi_j(t)$ for each subtype $j = 1, ..., J$ and model $\xi_j(t)$. For example, a restrictive model might make the assumption that the ratio is constant over time, $\xi_j(t) = \xi_j$ for each subtype. One would then estimate $\hat{\xi}_j = \frac{n_j}{n_+}$ from any representative sample of incident cases (an appropriate cohort or incident case-control study), with n_j and n_+ denoting the

number of subtype j cases and the total number of cases respectively.

An example of a more flexible model would be a piecewise constant function, parametrizing $\xi_j(t) = \xi_{jk}$ for selected intervals $t_k \leq t \leq t_{k+1}$. One could then estimate the observed fraction of subtype j cases among all cases, restricted to the age stratum defined by t_k, $\hat{\xi}_{jk} = \frac{n_{jk}}{n_{+k}}$, from a representative sample of incident cases with subtype and age information. One could choose to consider other parametric functions for $\xi_c(t)$; however, in most situations the constant or time-varying proportionality models should be sufficient.

A registry may also have some, but not all, the information necessary to determine subtype. This could be the case when one or more of the tumor characteristics that define the subtype have been recently recognized as important, and the registry has yet to record information on those characteristics. Revisiting the example of four breast cancer subtypes defined by the binary tumor characteristics ER and PR status, the registry may have recorded ER status but not PR status. This registry would provide marginal hazard rates by ER status, $\lambda_{m(ER+)}(t)$ and $\lambda_{m(ER-)}(t)$. In order to obtain the subtype-specific marginal hazard rates $\lambda_{m(ER+,PR+)}(t)$, $\lambda_{m(ER+,PR-)}(t)$, $\lambda_{m(ER-,PR+)}(t)$, $\lambda_{m(ER-,PR-)}(t)$, one could take the same approach but in this case weighting the ER-defined marginal hazards by estimates of the finer subtype distributions through a proportionality model fit with a supplemental data source. By noting that $\lambda_{m(ER+,PR+)}(t) + \lambda_{m(ER+,PR-)}(t) = \lambda_{m(ER+)}(t)$ and that $\lambda_{m(ER-,PR+)}(t) + \lambda_{m(ER-,PR-)}(t) = \lambda_{m(ER-)}(t)$, we see that is is only necessary to establish proportionality models for two of the four subtypes, say $\lambda_{m(ER+,PR+)}(t)$ and $\lambda_{m(ER-,PR+)}(t)$, as the remaining subtypes (in this case $\lambda_{m(ER+,PR-)}(t)$ and $\lambda_{m(ER-,PR-)}(t)$) are then defined.

2.4 Summary and Absolute Risk Predictions

In this chapter we have discussed a number of different ways to estimate the key components of a subtype-specific absolute risk model, depending on the characteristics of the available data sources. Figure 1 summarizes which data sources can be used to estimate each component, providing a road map for using different data sources in conjunction with one another to develop the model. For instance, we could fit an

Figure 2.1: Data Sources for Estimation When Available Incidence Rates Have Subtype Information

Goal:	To Fit the Model				
Needed:	1 Cox βc's	2 Baseline Hazard Function for Each Subtype C			3 Other Age-Specific Mortality
	↓	Age-Specific Incidence of Subtypes C	Distribution of Covariates		↓
			In the Population	In the Cases	
Cohort Study	✔	✔	✔	✔	✔
Case-Control Study	✔	✘	✔	✔	✘
Cancer Registry	✘	✔	✘	✘	✘
National Survey	✘	✘	✔	✘	✔

absolute risk model by estimating the β_j parameters and attributable risk (using the distribution of covariates among cases) from a case-control study and combining those estimates with subtype-specific marginal hazard rates from a registry and competing mortality rates from a national survey. We also discussed ways to handle registry data that is missing tumor subtype. Figure 2 adds these details to the overall picture of which data can be used to estimate each model component.

The option to build absolute risk models by integrating multiple data sources allows researchers to use the most appropriate choice available for each component of the model, drawing on the strengths of different types of data. When multiple

Figure 2.2: Data Sources for Estimation When Incidence Rates Are Missing Subtype Information

Goal:	To Fit the Model					
Needed:	1 Cox βc's	2 Baseline Hazard Function for Each Subtype C				3 Other Age-Specific Mortality
	↓	Age-Specific Incidence of Subtypes C		Distribution of Covariates		↓
		Overall Age-Specific Incidence	Distribution of Subtypes	In the Population	In the Cases	
Cohort Study	✔	✔	✔	✔	✔	✔
Case-Control Study	✔	✘	✔	✔	✔	✘
Cancer Registry	✘	✔	✘	✘	✘	✘
National Survey	✘	✘	✘	✔	✘	✔

sources of data are available to fit a given model component, careful consideration of characteristics such as the representativeness, sample size, sampling scheme, and missingness of the data should go into deciding which source is ultimately the best choice for estimating that component.

Typically once the subtype-specific absolute risk model is built, the goal is to use the model to learn about the distribution of absolute risk for the subtypes of interest in the population. This may mean looking at the proportion of the population that exceeds clinically relevent thresholds of risk or evaluating risk differences associated with a given risk factor for a particular subtype. Whatever the goal, in order to learn about the distribution of risk in the population one must first use the fitted model to predict risk for a set of covariate profiles that are representative of the population. This representative set of covariate profiles could come directly from a cohort study or from the controls in a population-based case-control study. The representative set of covariate profiles could also be simulated from the joint distribution of the covariates Z, perhaps estimated in a large national survey. In Chapter 4 we will demonstrate these ideas by fitting a subtype-specific absolute risk model to real data, making risk

predictions for breast cancer subtypes that are representative of the United States population, and analyzing the resulting risk distributions to evaluate the impact of risk factor modification on the distribution of risk for various subtypes. Before doing so, we will first address the statistical challenge of variance estimation, characterizing the uncertainty in risk predictions from an absolute risk model which is fit on multiple sources of data.

2.5 Variance Estimation

Characterizing the uncertainty in subtype-specific absolute risk estimates is critical to their use. An individual's risk is estimated by applying the fitted absolute risk model to the individual's set of covariates. Thus, the variance of the risk estimate is directly related to the variance of the parameters that comprise the fitted model. Variance estimation in this context is challenging because we must account for the fact that the risk estimate is a complex function of the various model parameters, which we have shown may be estimated from multiple different sources of data. A given data set may also be used to estimate more than one set of model parameters, in which case a valid variance estimation procedure must account for covariance between the estimates and how it ultimately affects the variance of the absolute risk estimate.

To construct a variance estimator that accounts for these features, we apply empirical process theory, basing the variance calculation on influence functions. The influence function approach has previously been applied to the problems of estimating the variance of attributable risk estimates from complex surveys and of estimating the variance of absolute risks for colorectal cancer (Graubard and Fears, 2005; Freedman et al., 2009b). We extend the approach to the setting of subtype-specific absolute risk models, handling influence functions for subtype-specific absolute risks,

the reparametrized multinomial model for subtypes, and models of the subtype distribution.

For illustrative purposes, suppose the model is parametrized by $\hat{\beta}_j, \hat{\lambda}_{0jq}$ for $j = 1, ..., J$ and $q = 1, ..., Q_j$ for each subtype j. Generally, a risk estimate for subtype j^*, denoted A_{j^*}, is simply a function of the parameter estimates $\hat{A}_{j^*} = f(\hat{\beta}_j, \hat{\lambda}_{0jq})$. Given this general form, we express the risk estimate as a linear function of the parameter estimates through a Taylor's expansion

$$\left(\hat{A}_{j^*} - A_{j*}\right) = \sum_{j=1}^{J} \left(\frac{\partial f(\beta_j, \lambda_{0jq})}{\partial \beta_j} \left(\hat{\beta}_j - \beta_j\right) + \sum_{q=1}^{Q_j} \frac{\partial f(\beta_j, \lambda_{0jq})}{\partial \lambda_{0jq}} \left(\hat{\lambda}_{0jq} - \lambda_{0jq}\right) \right) + o_p(1).$$

We then express each parameter estimate as a sum of independent identically distributed influence functions, which quantify the contribution of each data point to a given parameter estimate, with $\hat{\beta}_j - \beta_j = \frac{1}{N} \sum_{i=1}^{N} \psi_{ji}$ and $\hat{\lambda}_{0jq} - \lambda_{0jq} = \frac{1}{N} \sum_{i=1}^{N} \varphi_{jqi}$,

$$\begin{aligned}
\left(\hat{A}_{j^*} - A_{j^*}\right) =& \sum_{j=1}^{J} \left(\frac{\partial f(\beta_j, \lambda_{0jq})}{\partial \beta_j} \right) \left(\frac{1}{N} \sum_{i=1}^{N} \psi_{ji} \right) \\
&+ \sum_{j=1}^{J} \sum_{q=1}^{Q_j} \left(\frac{\partial f(\beta_j, \lambda_{0jq})}{\partial \lambda_{0jq}} \right) \left(\frac{1}{N} \sum_{i=1}^{N} \varphi_{jqi} \right) + o_p(1) \\
\sqrt{N}\left(\hat{A}_{j^*} - A_{j^*}\right) =& \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\sum_{j=1}^{J} \left(\left(\frac{\partial A_{j^*}}{\partial \beta_j} \right) (\psi_{ji}) + \sum_{q=1}^{Q_j} \left(\frac{\partial A_{j^*}}{\partial \lambda_{0jq}} \right) (\varphi_{jqi}) \right) \right] + o_p(1) \\
\sqrt{N}\left(\hat{A}_{j^*} - A_{j^*}\right) =& \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_i + o_p(1).
\end{aligned}$$

By using influence functions, we express the risk estimate as an sum of independent identically distributed contributions ϕ_i from each observation that was used to estimate the model parameters. Due to the central limit theorem, we know that the

variance of the risk estimate $Var\left[\hat{A}_{j^*}\right] = \frac{Var[\phi]}{N}$. With the proper formula for ϕ, we can estimate this variance empirically by computing the variance of ϕ_i for all individuals in the overall data. Again, this overall data may be comprised of observations from multiple data sources.

Thus far we have described a general strategy for constructing a variance estimator for risk estimates from a subtype-specific absolute risk model. However, the specific formula for ϕ will depend on how the model is parametrized and how those parameters are estimated. In this chapter we have provided many different options for modeling and model fitting, which depend on the chosen data sources. In the following, we will derive the formula of ϕ for one possible choice of model parametrization and fitting.

We present the mathematical details of variance calculation for the following model, which incorporates attributable risk and subtype ratio as described in 2.3.2:

$$A_{j^*} = \int_a^{a+\tau} \lambda_{0j^*}(t) e^{Z\beta_{j^*}} exp\left(-\int_a^t \left[\sum_{j=1}^{J} \lambda_{0j}(u) e^{Z\beta_j} + c(u)du\right]\right) dt$$

$$A_{j^*} = \int_a^{a+\tau} \left(1 - AR_{j^*}(t)\right) \xi_{j^*}(t) \lambda_{m+}(t) e^{Z\beta_{j^*}} \tag{2.21}$$

$$\cdot\, exp\left(-\int_a^t \left[\sum_{j=1}^{J} \left(1 - AR_j(t)\right) \xi_j(t) \lambda_{m+}(u) e^{Z\beta_j} + c(u)du\right]\right) dt.$$

In Chapter 4 we fit this model to real data, so the variance calculation presented here will reflect the characteristics of the data sources we use in the actual data analysis. For instance, we obtain estimates of competing mortality $c(t)$ from national survey data and overall marginal hazard $\lambda_{m+}(t)$ from a national cancer registry. These data sources have large enough sample sizes that the estimates $c(t)$ and $\lambda_{m+}(t)$ are essentially without variance, so in the variance calculation we treat these rates as

known. Estimation of the subtype-specific attributable risks $AR_j(t)$, the subtype-ratios $\xi_j(t)$, and the log hazard ratios β_j, all contribute variation to the variance of the risk estimate A_{j^*}.

The precise relationships depend on the model's parametrization, which we formalize as

$$AR_j(t) = AR_{jq} \text{ when } t_q \leq t \leq t_{q+1} \qquad \lambda_{m+}(t) = \lambda_k \text{ when } t_k = t$$

$$\xi_j(t) = \xi_{jp} \text{ when } t_p \leq t \leq t_{p+1} \qquad c(t) = c_l \text{ when } t_l = t.$$

We define the rates $c(t)$ and $\lambda_{m+}(t)$ by known parameters at integer ages t_l for $l = 1, ..., L$ and t_k for $k = 1, ..., K$ respectively. We define the functions $AR_j(t)$ and $\xi_j(t)$ by the subtype-specific scalar parameters AR_{jq} and ξ_{jp} in sequential time intervals with cutpoints t_q and t_p, allowing the functions to vary with time. Defining these functions with more and finer intervals allows more flexible modeling with time, but requires estimation of a greater number of parameters. The intervals should be selected to jointly cover the time interval where risk prediction is desired with as much precision as can be well-supported by the available data.

By expressing the model in terms of the parameters, we define the function $\hat{A}_{j^*} = f(\hat{\beta}_j, \hat{AR}_{jq}, \hat{\xi}_{jp})$ as

$$\hat{A}_{j^*} = \sum_{t=a}^{a+\tau} \left\{ \left(1 - \sum_q AR_{j^*q} I_{\{t_q \leq t \leq t_{q+1}\}} \right) \left(\sum_p \xi_{j^*p} I_{\{t_p \leq t \leq t_{p+1}\}} \right) \left(\sum_k \lambda_k I_{\{t_k = t\}} \right) e^{Z\beta_{j^*}} \right.$$

$$\cdot exp \left(- \sum_{u=a}^{t} \sum_{j=1}^{J} \left[\left(1 - \sum_q AR_{jq} I_{\{t_q \leq u \leq t_{q+1}\}} \right) \left(\sum_p \xi_{jp} I_{\{t_p \leq u \leq t_{p+1}\}} \right) \left(\sum_k \lambda_k I_{\{t_k = u\}} \right) e^{Z\beta_j} \right] \right.$$

$$\left. \left. - \sum_l c_l I_{\{t_l = u\}} \right) \right\}.$$

As before, the first step in constructing a variance estimator for this model is to apply

a Taylor's expansion to establish a linear relationship between $\hat{A}_{j^*}$ and the parameter estimates

$$\left(\hat{A}_{j^*} - A_{j^*}\right) = \sum_{j=1}^{J} \left(\frac{\partial f(\beta_j, AR_{jq}, \xi_{jp})}{\partial \beta_j}\right) \left(\hat{\beta}_j - \beta_j\right)$$

$$+ \sum_{j=1}^{J} \sum_{q=1}^{Q_j} \left(\frac{\partial f(\beta_j, AR_{jq}, \xi_{jp})}{\partial AR_{jq}}\right) \left(\hat{AR}_{jq} - AR_{jq}\right)$$

$$+ \sum_{j=1}^{J} \sum_{p=1}^{P_j} \left(\frac{\partial f(\beta_j, AR_{jq}, \xi_{jp})}{\partial \xi_{jp}}\right) \left(\hat{\xi}_{jp} - \xi_{jp}\right) + o_p(1).$$

Next we express the centered parameter estimates as $\left(\hat{\beta}_j - \beta_j\right) = \frac{1}{N} \sum_{i=1}^{N} \psi_{bji}$, $\left(\hat{AR}_{jq} - AR_{jq}\right) = \frac{1}{N} \sum_{i=1}^{N} \psi_{ajqi}$, and $\left(\hat{\xi}_{jp} - \xi_{jp}\right) = \frac{1}{N} \sum_{i=1}^{N} \psi_{cjpi}$. We incorporate the influence functions ψ_{bji}, ψ_{ajqi}, and ψ_{cjpi} and obtain

$$\sqrt{N}\left(\hat{A}_{j^*} - A_{j^*}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{j=1}^{J} \left[\left(\frac{\partial f(\beta_j, AR_{jq}, \xi_{jp})}{\partial \beta_j}\right) \psi_{bji} + \sum_{q=1}^{Q_j} \left(\frac{\partial f(\beta_j, AR_{jq}, \xi_{jp})}{\partial AR_{jq}}\right) \psi_{ajqi} \right.$$

$$\left. + \sum_{p=1}^{P_j} \left(\frac{\partial f(\beta_j, AR_{jq}, \xi_{jp})}{\partial \xi_{jp}}\right) \psi_{cjpi} \right] + o_p(1).$$

Thus, $\hat{A}_{j^*}$ is the sum of independent identically distributed contributions ϕ_i

$$\sqrt{N}\left(\hat{A}_{j^*} - A_{j^*}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_i + o_p(1) \text{ where} \tag{2.22}$$

$$\phi_i = \sum_{j=1}^{J} \left[\left(\frac{A_{j^*}}{\partial \beta_j}\right) \psi_{bji} + \sum_{q=1}^{Q_j} \left(\frac{A_{j^*}}{\partial AR_{jq}}\right) \psi_{aji} + \sum_{p=1}^{P_j} \left(\frac{A_{j^*}}{\partial \xi_{jp}}\right) \psi_{cji} \right].$$

To obtain the formula for ϕ_i, we must compute the derivative of A_{j^*} in terms of each parameter and obtain the formula for the influence functions ψ_{bji}, ψ_{ajqi}, and ψ_{cjpi}. The derivatives are provided in Appendix F and the derivation of the influence functions is given in Appendix G.

2.6 Simulations

To verify that the variance estimator given by equation (2.22) estimates the variance well, we evaluated its performance using simulations. We generated cohort data by simulating two binary covariates, Z_1 and Z_2 with probabilities 0.2 and 0.5. We then simulated event times, T_1 and T_2, for two subtypes, S_1 and S_2, from weibull distributions with shape parameters $K1$, $K2$ and scale parameters $\frac{L1}{exp(\frac{[Z_1,Z_2]\beta_1}{K1})}$, $\frac{L2}{exp(\frac{[Z_1,Z_2]\beta_2}{K2})}$ respectively, which induce the specified log hazard ratios $\beta_1 = [log(2), 0]'$ and $\beta_2 = [log(3), 0]'$. We selected parameters $K1 = 3.22$, $L1 = 209.1$, $K2 = 3.72$, and $L2 = 159.1$ such that the hazard rates of simulated event times approximately corresponded with breast cancer incidence rates in SEER. We also simulated potential censoring times T_c from a normal distribution with mean 70 and standard deviation 3.5. For each individual, the final event time $T = \min(T_1, T_2, T_c)$ with the subtype S assigned accordingly, taking subtype 0 if censoring time occurred first. Thus, the overall simulated data was comprised of variables $[Z_1, Z_2, T, S]$.

For 1000 simulated datasets of size 800000, we fit the absolute risk model and recorded absolute risk estimates for each of the two subtypes, and the corresponding variance estimates, for individuals with two different covariate profiles: $[Z_1 = 0, Z_2 = 0]$ and $[Z_1 = 1, Z_2 = 1]$. Based on our knowledge of the true data generating distributions, we computed the true values of absolute risk for each subtype. Table 2.4 presents the average absolute risk estimates for ages 50-70 along with the percent

bias on the scale of population risk per 1000 women.

Table 2.4: Average Estimates and Percent Bias of Absolute Risk in Ages 50-70 for Two Subtypes on the Scale of Risk in 1000 Women

	Covariate Profile	True Risk	Estimated Risk	% Bias
Subtype 1	$Z_1 = 0,\ Z_2 = 0$	19.92	19.91	-0.07
	$Z_1 = 1,\ Z_2 = 1$	38.05	38.07	0.05
Subtype 2	$Z_1 = 0,\ Z_2 = 0$	34.37	34.39	0.08
	$Z_1 = 1,\ Z_2 = 1$	98.32	98.51	0.19

The results show that our coded implementation of the model gives unbiased estimate of the subtype-specific absolute risks, with estimates deviating from the true values by less than one percent on average, in a non-systematic fashion. By presenting the risk estimates applied to a population of size 1000, we see that in a public health context the estimated number of predicted cancers would be off by substantially less than one person on average.

In Table 2.5 we examined the performance of the robust variance estimator by comparing the variance estimates to the "true" observed variation in risk estimates over the 1000 datasets. Table 2.5 presents the estimated standard deviations of the absolute risks in Table 2.4 as compared to the observed standard deviation, along with coverage probability. The results show that the robust variance estimator performs quite well, with coverage probability at the appropriate 95% level. These results demonstrate that the influence function based approach is a valid way to obtain variance estimates and is implemented correctly in our code.

Although the simulation results for the point estimates and robust variance estimates of the absolute risk clearly show that the implemented method is working well, to be thorough we also investigated the performance each estimated component in the

Table 2.5: Average Standard Error Estimates and Coverage Probabilities of Absolute Risk Estimates in Ages 50-70 for Two Subtypes on the Scale of Risk in 1000 Women

	Covariate Profile	Observed Standard Deviation	Estimated Standard Deviation	Coverage Probability
Subtype 1	$Z_1 = 0,\ Z_2 = 0$	0.223	0.219	0.944
	$Z_1 = 1,\ Z_2 = 1$	0.587	0.586	0.954
Subtype 2	$Z_1 = 0,\ Z_2 = 0$	0.287	0.284	0.954
	$Z_1 = 1,\ Z_2 = 1$	1.064	1.082	0.958

model: the log hazard ratios, the time-varying attributable risk, and the time-varying subtype ratio. In the simulation, we model $AR_j(t)$ and $\xi_j(t)$ by piecewise constant functions defined by four categories of time: 30-49, 50-58, 59-65, and 66-85.

Table 2.6 presents the average estimates and percent bias for each of the model components. The results show that across all model components the parameter estimates exhibit less than four percent bias on average, with less than one percent bias in the majority of the estimates. Similarly, Table 2.7 presents the average robust standard deviation estimates for each model component and the percent bias. Again, the results show that the robust variance estimator is performing well, with under five percent bias across all model components. Because the two subtype ratios for a given time interval add to one, one estimate is fully determined by the other and we expect them to have the same variance. This is the case in our simulation results.

Taken together, these simulation results empirically demonstrate the effectiveness of the robust variance estimator we constructed from the influence functions for each model parameter. This exercise also demonstrates that we implemented the method correctly in R and can feel confident applying our code to address practical problems using real data.

Table 2.6: Average Estimates and Percent Bias of Model Components for Absolute Risk Models for Two Subtypes, with Attributable Risk and Subtype Ratios Modeled by Piecewise Constant Functions Defined by Four Time Intervals

		Parameter	True Values	Average Estimates	% Bias
Log Hazard Ratios	Subtype 1	β_{z_1}	0.693	0.695	0.23
		β_{z_2}	0	-2.744E-04	–
	Subtype 2	β_{z_1}	1.099	1.100	0.13
		β_{z_2}	0	-2.337E-04	–
Attributable Risk, AR(t)	Subtype 1	$t \in [30, 49]$	0.165	0.165	-0.08
		$t \in [50, 58]$	0.161	0.161	0.01
		$t \in [59, 65]$	0.158	0.158	0.08
		$t \in [66, 85]$	0.150	0.154	3.19
	Subtype 2	$t \in [30, 49]$	0.283	0.282	-0.15
		$t \in [50, 58]$	0.278	0.278	0.05
		$t \in [59, 65]$	0.273	0.274	0.03
		$t \in [66, 85]$	0.260	0.267	2.60
Subtype Ratio, $\xi(t)$	Subtype 1	$t \in [30, 49]$	0.384	0.379	-1.33
		$t \in [50, 58]$	0.348	0.348	0.15
		$t \in [59, 65]$	0.334	0.333	-0.09
		$t \in [66, 85]$	0.314	0.322	2.47
	Subtype 2	$t \in [30, 49]$	0.616	0.621	0.83
		$t \in [50, 58]$	0.652	0.652	-0.08
		$t \in [59, 65]$	0.666	0.667	0.04
		$t \in [66, 85]$	0.686	0.678	-1.13

Table 2.7: Percent Bias of Robust Standard Deviation Estimates for Model Components in Two Absolute Risk Models, with Attributable Risk and Subtype Ratios Modeled by Piecewise Constant Functions Defined by Four Time Intervals

		Parameter	Observed Standard Deviation	Estimated Standard Deviation	% Bias
Log Hazard Ratios	Subtype 1	β_{z_1}	1.725E-02	1.708E-02	0.98
		β_{z_2}	1.611E-02	1.584E-02	1.64
	Subtype 2	β_{z_1}	1.348E-02	1.355E-02	-0.50
		β_{z_2}	1.300E-02	1.292E-02	0.59
Attributable Risk, AR(t)	Subtype 1	$t \in [30, 49]$	8.085E-03	8.033E-03	0.64
		$t \in [50, 58]$	8.267E-03	8.081E-03	2.25
		$t \in [59, 65]$	8.221E-03	8.221E-03	0.01
		$t \in [66, 85]$	8.414E-03	8.345E-03	0.81
	Subtype 2	$t \in [30, 49]$	6.189E-03	6.040E-03	2.41
		$t \in [50, 58]$	5.915E-03	6.005E-03	-1.52
		$t \in [59, 65]$	6.208E-03	6.214E-03	-0.11
		$t \in [66, 85]$	6.378E-03	6.422E-03	-0.69
Subtype Ratio, $\xi(t)$	Subtype 1	$t \in [30, 49]$	3.681E-03	3.662E-03	0.51
		$t \in [50, 58]$	3.755E-03	3.603E-03	4.04
		$t \in [59, 65]$	3.531E-03	3.457E-03	2.10
		$t \in [66, 85]$	3.223E-03	3.360E-03	-4.24
	Subtype 2	$t \in [30, 49]$	3.681E-03	3.662E-03	0.51
		$t \in [50, 58]$	3.755E-03	3.603E-03	4.04
		$t \in [59, 65]$	3.531E-03	3.457E-03	2.10
		$t \in [66, 85]$	3.223E-03	3.360E-03	-4.24

2.7 Appendix A: Equivalence Between Two Partial Likelihood Estimators

Beginning with the score function for the unconditional partial likelihood, we assign non-parametric estimates for the hazard at time t and show that the resulting score function is the same if we assume that there are no tied failure times. In doing so, we show that in such a situation the unconditional partial likelihood estimator $\tilde{\beta_{j^*}}$ is equivalent to the conditional partial likelihood estimator $\hat{\beta_{j^*}}$

$$S(\beta_{j^*}) = \frac{\partial log PL'(\beta_1, ..., \beta_J)}{\partial \beta_{j^*}} = \sum_{(i)} \left[\frac{\lambda_{0j^*}(t_{(i)}) z_i e^{\beta_{j^*} z_i}}{\sum_{j=1}^{J} \lambda_{0j}(t_{(i)}) e^{\beta_j z_i}} - \frac{\sum_{l \in R_{t_{(i)}}} \lambda_{0j^*}(t_{(i)}) z_l e^{\beta_{j^*} z_l}}{\sum_{l \in R_{t_{(i)}}} \sum_{j=1}^{J} \lambda_{0j}(t_{(i)}) e^{\beta_j z_l}} \right].$$

In Appendix B the non-parametric estimates defining the baseline hazard function at time t_{jq} are shown to be $\hat{\lambda_{jq}} = \frac{d_j(t_{jq})}{\sum_{v \in R(t_{jq})} e^{\beta_j z_i}}$. Plugging these in, we find

$$S(\beta_{j^*}) = \sum_{(i)} \left[\frac{\frac{d_{j^*}(t(i))}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}} z_i e^{\beta_{j^*} z_i}}{\sum_{j=1}^{J} \frac{d_j(t(i))}{\sum_{v \in R_{t_{(i)}}} e^{\beta_j Z_v}} e^{\beta_j z_i}} - \frac{\sum_{l \in R_{t_{(i)}}} \frac{d_{j^*}(t(i))}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}} z_l e^{\beta_{j^*} z_l}}{\sum_{l \in R_{t_{(i)}}} \sum_{j=1}^{J} \frac{d_j(t(i))}{\sum_{v \in R_{t_{(i)}}} e^{\beta_j Z_v}} e^{\beta_j z_l}} \right].$$

Parts of the denominator in the second term cancel out immediately, leaving

$$S(\beta_{j^*}) = \sum_{(i)} \left[\frac{\frac{d_{j^*}(t_{(i)})}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}} z_i e^{\beta_{j^*} z_i}}{\sum_{j=1}^{J} \frac{d_j(t_{(i)})}{\sum_{v \in R_{t_{(i)}}} e^{\beta_j Z_v}} e^{\beta_j z_i}} - \frac{\sum_{l \in R_{t_{(i)}}} \frac{d_{j^*}(t_{(i)})}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}} z_l e^{\beta_{j^*} z_l}}{\sum_{j=1}^{J} d_j(t(i))} \right].$$

If we assume that there are no tied failure times, then each $d_j(t_{(i)})$ is 0 or 1 based on whether the single observed event at that time was of subtype j. This means that $\sum_{j=1}^{J} d_j(t_{(i)}) = 1$. Moreover, when the single event at a given time is not of subtype j^* then $d_j^*(t_{(i)}) = 0$, and thus the contribution to the overall sum is zero at those times

$$S(\beta_{j^*}) = \sum_{(i):S_i=j^*} \left[\frac{\frac{1}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}} z_i e^{\beta_{j^*} z_i}}{\sum_{j=1}^{J} \frac{d_j(t_{(i)})}{\sum_{v \in R_{t_{(i)}}} e^{\beta_j Z_v}} e^{\beta_j z_i}} - \frac{\sum_{l \in R_{t_{(i)}}} \frac{1}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}} z_l e^{\beta_{j^*} z_l}}{1} \right].$$

The only non-zero contribution in the denominator of the first term occurs when $j = j^*$, in which case $d_j^*(t_{(i)}) = 1$ and the rest equal 0, giving

$$S(\beta_{j^*}) = \sum_{(i):S_i=j^*} \left[\frac{\frac{z_i e^{\beta_{j^*} z_i}}{\sum_{v \in R_{t_{(i)}}} e^{\beta_j Z_v}}}{\frac{e^{\beta_{j^*} z_i}}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}}} - \sum_{l \in R_{t_{(i)}}} \frac{z_l e^{\beta_{j^*} z_l}}{\sum_{v \in R_{t_{(i)}}} e^{\beta_{j^*} Z_v}} \right]$$

$$S(\beta_{j^*}) = \sum_{(i):S_i=j^*} \left[z_i - \frac{\sum_{l \in R_{t_{(i)}}} z_l e^{\beta_{j^*} z_l}}{\sum_{v \in R_{t_{(i)}}} e^{\beta_j Z_v}} \right].$$

This is the score equation for the partial likelihood conditional on subtype.

2.8 Appendix B: Deriving the Subtype-Specific Baseline Hazard Estimates

In the following equations, we derive the maximum likelihood estimator for parameters λ_{jq} which non-parametrically define the subtype-specific baseline hazard functions λ_{0j}. This likelihood formulation assumes that censoring is independent of the

subtype outcomes, given the covariates Z

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_j(t_i|z_i) S(t_i|z_i) \right) \left(\prod_{i:S_i=0} S(t_i|z_i) \right)$$

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_j(t_i|z_i) \right) \left(\prod_{i=1}^{N} S(t_i|z_i) \right)$$

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_j(t_i|z_i) \right) \left(\prod_{i=1}^{N} exp\left\{-\Lambda(t_i|z_i)\right\} \right)$$

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_j(t_i|z_i) \right) \left(\prod_{i=1}^{N} exp\left\{ - \int_0^{t_i} \lambda(u|z_i) du \right\} \right)$$

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_j(t_i|z_i) \right) \left(\prod_{i=1}^{N} exp\left\{ - \int_0^{t_i} \sum_{j=1}^{J} \lambda_j(u|z_i) du \right\} \right)$$

We then incorporate the proportional hazards model, $\lambda_j(t|Z) = \lambda_{0j}(t) e^{\beta_j Z}$, into the likelihood

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_{0j}(t_i) e^{\beta_j z_i} \right) \left(\prod_{i=1}^{N} exp\left\{ - \int_0^{t_i} \sum_{j=1}^{J} \lambda_{0j}(u) e^{\beta_j z_i} du \right\} \right).$$

Recall the non-parametric characterization of $\lambda_{0j}(t)$ and $\Lambda_{0j}(t)$ in terms of constant parameters λ_{jq}

$$\lambda_{0j}(t) = \left\{ \begin{array}{ll} \lambda_{jq} & \text{for } t_{jq} = t \\ 0 & \text{else} \end{array} \right\} = \sum_{q=1}^{Q} \lambda_{jq} I\left\{t_{jq} = t\right\}, \quad \Lambda_{0j}(t) = \sum_{m:t_{jm}<t} \lambda_{jm}.$$

We express the likelihood in terms of this characterization as

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} \left(\sum_{q=1}^{Q} I\left\{S_i = j\right\} \lambda_{jq} I\left\{t_{jq} = t_i\right\} \right) e^{\beta_j z_i} \right) \cdot \left(\prod_{i=1}^{N} exp\left\{ - \sum_{j=1}^{J} \left(\sum_{m:t_{jm} < t_i} \lambda_{jm} \right) e^{\beta_j z_i} \right\} \right)$$

$$log(L) = \sum_{i:S_i \geq 1} \sum_{j=1}^{J} \left(log(\sum_{q=1}^{Q} I\left\{S_i = j\right\} \lambda_{jq} I\left\{t_{jq} = t_i\right\}) + \beta_j z_i \right) - \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\sum_{m:t_{jm} < t_i} \lambda_{jm} \right) e^{\beta_j z_i}.$$

We subsequently derive the score function for a specific λ_{j^*q*} and solve for $\hat{\lambda}_{j^*q*}$

$$S(\lambda_{j^*q*}) = \frac{\partial log L}{\partial \lambda_{j^*q*}} = \sum_{i:S_i \geq 1} \frac{I\left\{S_i = j^*\right\} I\left\{t_{j^*q^*} = t_i\right\}}{\sum_{q=1}^{Q} I\left\{S_i = j^*\right\} \lambda_{j^*q} I\left\{t_{j^*q} = t_i\right\}} - \sum_{i=1}^{N} I\left\{t_{j^*_{q*}} < t_i\right\} e^{\beta_{j^*} z_i}$$

$$S(\lambda_{j^*q*}) = \sum_{i:S_i = j^*} \frac{I\left\{t_{j^*q^*} = t_i\right\}}{\lambda_{j^*q*}} - \sum_{i:t_{j^*q^*} < t_i}^{N} e^{\beta_{j^*} z_i} = 0.$$

Thus, the maximum likelihood estimator for a given parameter λ_{j^*q*} is

$$\hat{\lambda_{j^*q*}} = \frac{\sum_{i:S_i = j^*} I\left\{t_{j^*q^*} = t_i\right\}}{\sum_{i:t_{j^*q^*} < t_i}^{N} e^{\beta_{j^*} z_i}} = \frac{d_{j^*}(t_{j^*q^*})}{\sum_{R(t_{j^*q^*})} e^{\beta_{j^*} z_i}}.$$

2.9 Appendix C: Deriving the Subtype-Specific Baseline Hazard Estimates, Drawing Strength Across Subtypes

Here we derive the non-parametric estimator of the referent baseline hazard function. We start with the likeilhood constructed in Appendix B

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_j(t_i|z_i) \right) \left(\prod_{i=1}^{N} exp\left\{ - \int_0^{t_i} \sum_{j=1}^{J} \lambda_j(u|z_i) du \right\} \right).$$

We incorporate the new formulation of the proportional hazards model $\lambda_j(t) = \lambda_{01}(t) h_j(\theta_j, t) e^{\beta_j z_i}$ based on a proportional relationship between the baseline hazard functions of different subtypes

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_{01}(t) h_j(\theta_j, t) e^{\beta_j z_i} \right) \cdot \left(\prod_{i=1}^{N} exp\left\{ - \int_0^{t_i} \sum_{j=1}^{J} \lambda_{01}(t) h_j(\theta_j, t) e^{\beta_j z_i} du \right\} \right).$$

Recall that we define the reference baseline hazard function $\lambda_{01}(t)$ non-parametrically by parameters $\lambda_1, ..., \lambda_Q$ at each observed event time $t_1, ..., t_Q$, with zero hazard in between. The cumulative baseline hazard function is the sum of each parameter preceding time t

$$\lambda_{01}(t) = \left\{ \begin{array}{ll} \lambda_q & \text{for } t_q = t \\ 0 & \text{else} \end{array} \right\} = \sum_{q=1}^{Q} \lambda_q I\left\{t_q = t\right\}, \quad \Lambda_{01}(t) = \sum_{m:t_m < t} \lambda_m.$$

We express the likelihood in terms of the parameters $\lambda_1, ..., \lambda_Q$ as

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} \sum_{q=1}^{Q} I\left\{S_i = j\right\} \lambda_q I\left\{t_q = t\right\} h_j(\theta_j, t) e^{\beta_j z_i} \right) \cdot \left(\prod_{i=1}^{N} exp\left\{ -\sum_{j=1}^{J} \sum_{m:t_m < t_i} \lambda_m h_j(\theta_j, t_m) e^{\beta_j z_i} \right\} \right).$$

$$log(L) = \sum_{i:S_i \geq 1} \sum_{j=1}^{J} log\left\{ \sum_{q=1}^{Q} I\left\{S_i = j\right\} \lambda_q I\left\{t_q = t_i\right\} \right\} + log(h_j(\theta_j, t_i)) + \beta_j z_i - \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{m:t_m < t_i} \lambda_m h_j(\theta_j, t_m) e^{\beta_j z_i}$$

To obtain the profile likelihood estimate of a specific λ_{q*}, we treat $h_j(\theta_j, t_i)$ and β_j as fixed and solve the score function for λ_q

$$\frac{\partial log(L)}{\partial \lambda_{q*}} = \sum_{i:S_i \geq 1} \frac{I\left\{t_{q*} = t_i\right\}}{\sum_{q=1} I\left\{S_i = j^*\right\} \lambda_q I\left\{t_{j^*} = t_i\right\}} - \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{m:t_m < t_i} h_j(\theta_j, t_j) e^{\beta_j z_i}$$

$$\frac{\partial log(L)}{\partial \lambda_{q*}} = \sum_{i:S_i \geq 1} \frac{I\left\{t_{q*} = t_i\right\}}{\lambda_{q*}} - \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{m:t_m < t_i} h_j(\theta_j, t_j) e^{\beta_j z_i} = 0.$$

Thus, the profile likelihood estimator for a given referent baseline hazard parameter parameter λ_{q*} is

$$\hat{\lambda}_{q*} = \frac{\sum_{i:S_i \geq 1} I\left\{t_{q*} = t_i\right\}}{\sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{m:t_m < t_i} h_j(\theta_j, t_j) e^{\beta_j z_i}} = \frac{d_{(t_{q*})}}{\sum_{i \in R_{(t_{q*})}} \sum_{j=1}^{J} h_j(\theta_j, t_j) e^{\beta_j z_i}}.$$

2.10 Appendix D: Estimating the Proportionality Model Relating Subtype Baseline Hazard Functions

To estimate the general function $h_j(\theta_j, t_i)$ for a particular subtype, we will obtain the score function from the profile likelihood of θ_j. We begin with the likelihood from Appendix C, calculate the log likelihood, and take the derivative with respect to a particular θ_{j^*}

$$L = \left(\prod_{i:S_i \geq 1} \prod_{j=1}^{J} I\left\{S_i = j\right\} \lambda_{01}(t_i) h_j(\theta_j, t_i) e^{\beta_j z_i} \right) \cdot \left(\prod_{i=1}^{N} exp\left\{ - \int_0^{t_i} \sum_{j=1}^{J} \lambda_{01}(u) h_j(\theta_j, u) e^{\beta_j z_i} du \right\} \right)$$

$$log(L) = \left(\sum_{i:S_i \geq 1} \sum_{j=1}^{J} I\left\{S_i = j\right\} \left[log(\lambda_{01}(t_i)) + log(h_j(\theta_j, t_i)) + \beta_j z_i \right] \right) \cdot + \left(\sum_{i=1}^{N} - \int_0^{t_i} \sum_{j=1}^{J} \lambda_{01}(u) h_j(\theta_j, u) e^{\beta_j z_i} du \right)$$

$$\frac{\partial log(L)}{\partial \theta_{j^*}} = \left(\sum_{i:S_i = j^*} \frac{\left(\frac{\partial h_{j^*}(\theta_{j^*}, t_i)}{\partial \theta_{j^*}} \right)}{h_{j^*}(\theta_{j^*}, t_i)} \right) + \left(\sum_{i=1}^{N} - \int_0^{t_i} \lambda_{01}(u) \frac{\partial h_{j^*}(\theta_{j^*}, u)}{\partial \theta_{j^*}} e^{\beta_{j^*} z_i} du \right).$$

Thus the estimate $\hat{\theta}_{j^*}$ is the value θ_{j^*} that solves

$$\frac{\partial log(L)}{\partial \theta_{j^*}} = \left(\sum_{i:S_i = j^*} \frac{h'_{j^*}(\theta_{j^*}, t_i)}{h_{j^*}(\theta_{j^*}, t_i)} \right) + \left(\sum_{i=1}^{N} - \int_0^{t_i} \lambda_{01}(u) h'_{j^*}(\theta_{j^*}, u) e^{\beta_{j^*} z_i} du \right) = 0.$$

In general, estimation of θ_{j^*} depends on the referent baseline hazard function $\lambda_{01}(t_i)$ and β_{j^*}. Thus, obtaining final estimates necessitates iteratively solving the estimating equations for all the parameters.

In the case of the simple proportionality model where $h_j(\theta_j, t_i) = \theta_j$, and correspondingly $h'_j(\theta_j, t) = 1$, the general equation simplifies and can be solved analytically

$$\frac{\partial log(L)}{\partial \theta_{j^*}} = \left(\sum_{i:S_i=j^*} \frac{1}{\theta_{j^*}} \right) + \left(\sum_{i=1}^{N} - \int_0^{t_i} \lambda_{01}(u) e^{\beta_{j^*} z_i} du \right) = 0$$

$$\hat{\theta}_{j^*} = \frac{d_{j^*}}{\sum_{i=1}^{N} - \int_0^{t_i} \lambda_{01}(u) e^{\beta_{j^*} z_i} du},$$

where d_{j^*} is the number of observed events of subtype j^* irrespective of the time that those events occurred. This is not particularly surprising due to the fact that the model $h_j(\theta_j, t_i) = \theta_j$ is not time-varying.

2.11 Appendix E: Influence Function for $\frac{\bar{x}}{\bar{y}}$

Suppose x_i and y_i are independent and identically distributed where $x \sim F_0$, $\mu_x = E[x]$ and $y \sim G_0$, $\mu_y = E[y]$. Our goal is to derive the influence function ψ_i associated with the estimator $\frac{\bar{x}}{\bar{y}}$. To do this, we first express the estimator $\frac{\bar{x}}{\bar{y}}$ as a function of the empirical distributions F_n and G_n, and the parameter $\frac{\mu_x}{\mu_y}$ as a function of the true distributions F_0 and G_0

$$\left(\frac{\bar{x}}{\bar{y}} - \frac{\mu_x}{\mu_y} \right) = \left(\frac{\int x \; d\{F_n\}}{\int y \; d\{G_n\}} - \frac{\int x \; d\{F_0\}}{\int y \; d\{G_0\}} \right) = \phi \begin{pmatrix} F_n \\ G_n \end{pmatrix} - \phi \begin{pmatrix} F_0 \\ G_0 \end{pmatrix} \cong \phi'_{(F_0,G_0)} \begin{pmatrix} F_n - F_0 \\ G_n - G_0 \end{pmatrix}.$$

The difference $\left(\frac{\bar{x}}{\bar{y}} - \frac{\mu_x}{\mu_y} \right)$ can be expressed as the derivative of the function ϕ that maps the distribution functions to the estimate and parameter. The empirical distributions

F_n and G_n are defined by n jump points with one parameter at each data point as

$$F_n = \frac{1}{n}\sum_{i=1}^{N}\delta_{x_i}(x) \text{ and } G_n = \frac{1}{n}\sum_{i=1}^{N}\delta_{y_i}(y), \text{ with } \delta_{x_i}(x) = \left\{ \begin{smallmatrix} 0,\ x_i>x \\ 1,\ x_i\leq x \end{smallmatrix} \right\},\ \delta_{y_i}(y) = \left\{ \begin{smallmatrix} 0,\ y_i>y \\ 1,\ y_i\leq y \end{smallmatrix} \right\}.$$

Given this definition, and the fact that ϕ' is a linear map,

$$\phi'_{(F_0,G_0)} \left(\begin{smallmatrix} F_n - F_0 \\ G_n - G_0 \end{smallmatrix} \right) = \phi'_{(F_0,G_0)} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{N}\delta_{x_i}(x) - F_0 \\ \frac{1}{n}\sum_{i=1}^{N}\delta_{y_i}(y) - G_0 \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{N}\phi'_{(F_0,G_0)} \left(\begin{smallmatrix} \delta_{x_i}(x) - F_0 \\ \delta_{y_i}(y) - G_0 \end{smallmatrix} \right).$$

Thus, we can express $\left(\frac{\bar{x}}{\bar{y}} - \frac{\mu_x}{\mu_y}\right)$ as a sum of indpendent identically distributed influence functions $\psi_i = \phi'_{(F_0,G_0)} \left(\begin{smallmatrix} \delta_{x_i}(x) - F_0 \\ \delta_{y_i}(y) - G_0 \end{smallmatrix} \right)$. We take this derivative in the direction of the true distributions by taking the derivative of ϕ along a "line" in the space of distributions, defined as a convex combination of the empirical and true distributions according to parameter t

$$\psi_i = \phi'_{(F_0,G_0)} \left(\begin{smallmatrix} \delta_{x_i}(x) - F_0 \\ \delta_{y_i}(y) - G_0 \end{smallmatrix} \right) = \frac{d}{dt}|_{t=0}\ \phi \left(\begin{smallmatrix} (1-t)\ F_0 + t\ \delta_{x_i}(x) \\ (1-t)\ G_0 + t\ \delta_{y_i}(y) \end{smallmatrix} \right)$$

$$\psi_i = \frac{d}{dt}|_{t=0} \frac{\int x\ d\{(1-t)\ F_0 + t\ \delta_{x_i}(x)\}}{\int y\ d\{(1-t)\ G_0 + t\ \delta_{y_i}(y)\}}$$

$$\psi_i = \frac{d}{dt}|_{t=0} \frac{(1-t)\int x\ d\{F_0\} + t\int x\ d\{\delta_{x_i}(x)\}}{(1-t)\int y\ d\{G_0\} + t\int y\ d\{\delta_{y_i}(y)\}}.$$

We simplify this expression by recognizing the relationships

$$\int x\ d\{F_0\} = \mu_x \qquad \int x\ d\{\delta_{x_i}(x)\} = x_i$$

$$\int y\ d\{G_0\} = \mu_y \qquad \int y\ d\{\delta_{y_i}(y)\} = y_i.$$

We then take the derivative with respect to t and evaluate at $t = 0$ to obtain the formula for the influence function

$$\psi_i = \frac{d}{dt}\big|_{t=0} \frac{(1-t)\mu_x+(t)x_i}{(1-t)\mu_y+(t)y_i}$$

$$\psi_i = \frac{(x_i-\mu_x)((1-t)\mu_y+td)-(y_i-\mu_y)((1-t)\mu_x+tb)}{((1-t)\mu_y+td)^2}$$

$$\psi_i = \frac{(x_i-\mu_x)\mu_y+(\mu_y-y_i)\mu_x}{\mu_y^2}$$

$$\psi_i = \frac{bc-ad}{\mu_y^2}$$

$$\psi_i = \frac{x_i\mu_y-y_i\mu_x}{\mu_y^2}.$$

Thus, the influence function for estimators of the form $\frac{\bar{x}}{\bar{y}}$ is $\psi_i = \frac{x_iE[y]-y_iE[x]}{E[y]^2}$.

2.12 Appendix F: Derivatives for the Variance Calculation

We solve for the derivatives $\frac{A_{j^*}}{\partial\beta_j}$, $\frac{A_{j^*}}{\partial AR_{jq}}$, and $\frac{A_{j^*}}{\partial\xi_{jp}}$ through a combination of straightforward calculus and careful bookkeeping. For ease of calculation, we work with an expression of A_{j^*} where all model terms are inside the exponential

$$A_{j^*} = \sum_{t=a}^{a+\tau} exp\left\{ log\left(1 - \sum_q AR_{j^*q} I_{\{t_q \le t \le t_{q+1}\}}\right) + log\left(\sum_p \xi_{j^*p} I_{\{t_p \le t \le t_{p+1}\}}\right) + log\left(\sum_k \lambda_k I_{\{t_k = t\}}\right) \right.$$
$$+ Z\beta_{j^*} - \sum_{u=a}^{t}\sum_{j=1}^{J}\left[\left(1 - \sum_q AR_{jq} I_{\{t_q \le u \le t_{q+1}\}}\right)\left(\sum_p \xi_{jp} I_{\{t_p \le u \le t_{p+1}\}}\right)\left(\sum_k \lambda_k I_{\{t_k = u\}}\right) e^{Z\beta_j}\right]$$
$$\left. - \sum_l c_l I_{\{t_l = u\}} \right\}.$$

Generally, $A_{j^*} = \sum_{t=a}^{a+\tau} exp\left(h\left(\beta_j, AR_{jq}, \xi_{jp}\right)\right)$. Thus, the derivatives needed for the calculation of ϕ are

$$\frac{A_{j^*}}{\partial \beta_{j'}} = \sum_{t=a}^{a+\tau} exp\left(h\left(\beta_j, AR_{jq}, \xi_{jp}\right)\right) Z \left\{ I_{\{j^*=j'\}} - e^{Z\beta_{j'}} \sum_{u=a}^{t} \left[\left(1 - \sum_q AR_{j'q} I_{\{t_q \le u \le t_{q+1}\}} \right) \cdot \left(\sum_p \xi_{j'p} I_{\{t_p \le u \le t_{p+1}\}} \right) \left(\sum_k \lambda_k I_{\{t_k = u\}} \right) \right] \right\},$$

$$\frac{A_{j^*}}{\partial AR_{j'q'}} = \sum_{t=a}^{a+\tau} exp\left(h\left(\beta_j, AR_{jq}, \xi_{jp}\right)\right) \left\{ -\frac{I_{\left\{t_{q'} \le t \le t_{q'+1}\right\}}}{1 - AR_{j'q'}} + \sum_{u=a}^{t} \left[I_{\left\{t_{q'} \le u \le t_{q'+1}\right\}} \cdot \left(\sum_p \xi_{j'p} I_{\{t_p \le u \le t_{p+1}\}} \right) \left(\sum_k \lambda_k I_{\{t_k = u\}} \right) e^{Z\beta_{j'}} \right] \right\},$$

$$\frac{A_{j^*}}{\partial \xi_{j'p'}} = \sum_{t=a}^{a+\tau} exp\left(h\left(\beta_j, AR_{jq}, \xi_{jp}\right)\right) \left\{ \frac{I_{\left\{t_{p'} \le t \le t_{p'+1}\right\}}}{\xi_{j'p'}} - \sum_{u=a}^{t} \left[I_{\left\{t_{p'} \le t \le t_{p'+1}\right\}} \cdot \left(1 - \sum_q AR_{j'q} I_{\{t_q \le u \le t_{q+1}\}} \right) \left(\sum_k \lambda_k I_{\{t_k = u\}} \right) e^{Z\beta_{j'}} \right] \right\}.$$

2.13 Appendix G: Influence Functions for the Variance Calculation

The final elements needed for the variance calculation are the influence functions, the independent identically distributed contributions of each observation to the parameters, such that $\left(\hat{\beta}_j - \beta_j\right) = \frac{1}{N} \sum_{i=1}^{N} \psi_{bji}$, $\left(\hat{AR}_{jq} - AR_{jq}\right) = \frac{1}{N} \sum_{i=1}^{N} \psi_{ajqi}$, and

$\left(\hat{\xi}_{jp} - \xi_{jp}\right) = \frac{1}{N}\sum_{i=1}^{N} \psi_{cjpi}$. The form of these influence functions depends on how the parameters are estimated.

When we ultimately fit the model in Chapter 4, we estimate the Cox parameters β_j using the multinomial likelihood method applied to cohort data. In Section 2.2.3 we derived the efficient influence function associated with the multinomial maximum likelihood estimator, accounting for the presence of nuisance parameters. With expectations replaced by empirical means, those formulas can be used to compute ψ_{bji} for each observation in the cohort data used to estimate β_j. For all observations that were in some way used in model fitting, but that did not contribute to estimation of the $\hat{\beta}_j$'s, the influences $\psi_{bji} = 0$.

The influence function ψ_{cjpi} should reflect the independent and identically distributed contribution of each observation to

$$\hat{\xi}_{jp} = \frac{\sum_{i=1}^{N} I_{\{S_i=j\}} I_{\{t_p \le t_i \le t_{p+1}\}}}{\sum_{i=1}^{N} I_{\{t_p \le t_i \le t_{p+1}\}}},$$

the empirical estimate of the proportion tumors in time interval $[t_p, t_{p+1}]$ that are subtype j. In Appendix E, we derive the influence function for estimators of this general form: $\frac{\bar{x}}{\bar{y}}$. Applying the resulting influence function to $\hat{\xi}_{jp}$, we can show that

$$\left(\hat{\xi}_{jp} - \xi_{jp}\right) = \frac{1}{N}\sum_{i=1}^{N} \frac{E\left[I_{\{t_p \le t \le t_{p+1}\}}\right] I_{\{S_i=j\}} I_{\{t_p \le t_i \le t_{p+1}\}} - E\left[I_{\{S=j\}} I_{\{t_p \le t \le t_{p+1}\}}\right] I_{\{t_p \le t_i \le t_{p+1}\}}}{E\left[I_{\{t_p \le t \le t_{p+1}\}}\right]^2}.$$

In the actual variance estimate, we approximate the expectations with empirical

means such that the influence function that we compute for each data point is

$$\psi_{cjpi} = \frac{\left[\frac{1}{N}\sum_{i=1}^{N} I_{\{t_p \leq t_i \leq t_{p+1}\}}\right] I_{\{S_i=j\}} I_{\{t_p \leq t_i \leq t_{p+1}\}} - \left[\frac{1}{N}\sum_{i=1}^{N} I_{\{S_i=j\}} I_{\{t_p \leq t_i \leq t_{p+1}\}}\right] I_{\{t_p \leq t_i \leq t_{p+1}\}}}{\left[\frac{1}{N}\sum_{i=1}^{N} I_{\{t_p \leq t_i \leq t_{p+1}\}}\right]^2}.$$

As before, for all observations that contributed to model fitting but were not used to estimate $\hat{\xi}_{jp}$, the influences $\psi_{cjpi} = 0$.

Finally, we derive the influence function ψ_{ajqi} associated with $\hat{AR}_{jq}$. In Chapter 4, we estimate $\hat{AR}_{jq}$ in time strata defined by $[t_q, t_{q+1}]$ using the Bruzzi formula described in Section 2.3.2 applied to incident cases from cohort data

$$\hat{AR}_{jq} = 1 - \left(\frac{1}{\sum_{i=1}^{N} I_{\{S_i=j\}} I_{\{t_q \leq t_i \leq t_{q+1}\}}}\right)\left(\sum_{i=1}^{N} \frac{I_{\{S_i=j\}} I_{\{t_q \leq t_i \leq t_{q+1}\}}}{e^{\hat{\beta}_j Z_i}}\right)$$

$$\hat{AR}_{jq} = 1 - \left(\frac{1}{N \cdot \hat{P}_{jq}}\right)\left(\sum_{i=1}^{N} \frac{I_{\{S_i=j\}} I_{\{t_q \leq t_i \leq t_{q+1}\}}}{e^{\hat{\beta}_j Z_i}}\right) \text{ with } \hat{P}_{jq} = \frac{1}{N}\sum_{i=1}^{N} I_{\{S_i=j\}} I_{\{t_q \leq t_i \leq t_{q+1}\}}.$$

The variance of $\hat{AR}_{jq}$ depends on the variance of $\hat{\beta}_j$, $\hat{P}_{jq}$, and the empirical variance in the distribution of the indicators. To take all these sources of variation into account in the influence function of $\hat{AR}_{jq}$, we again use a Taylor's approximation and influence

functions for each component

$$\left(\hat{AR}_{jq} - AR_{jq}\right) = \left(\frac{\partial AR_{jq}}{\partial \beta_j}\right)\left(\hat{\beta}_j - \beta_j\right) + \left(\frac{\partial AR_{jq}}{\partial P_{jq}}\right)\left(\hat{P}_{jq} - P_{jq}\right)$$

$$+ \phi'_{F_0}\left(F_n - F_0\right) + o_p(1)$$

$$\sqrt{N}\left(\hat{AR}_{jq} - AR_{jq}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\left(\frac{\partial AR_{jq}}{\partial \beta_j}\right)\psi_{bji} + \left(\frac{\partial AR_{jq}}{\partial P_{jq}}\right)\psi_{pjqi} + \psi_{eji}\right] + o_p(1).$$

Let F_0 denote the joint distribution of $S, T,$ and Z

$$AR = 1 - \frac{1}{N \cdot P_{jq}}\left(\int \frac{I_{\{S=j\}}\ I_{\{t_q \le t \le t_{q+1}\}}}{e^{\beta_j Z}}\ d\left\{F_0\right\}\right).$$

The necessary derivatives are

$$\left(\frac{\partial AR_{jq}}{\partial \beta_j}\right) = \frac{1}{N \cdot P_{jq}}\left(\int \frac{Z\ I_{\{S=j\}}\ I_{\{t_q \le t \le t_{q+1}\}}}{e^{\beta_j Z}}\ d\left\{F_0\right\}\right) = \frac{1}{N \cdot P_{jq}} E\left[\frac{Z\ I_{\{S=j\}}\ I_{\{t_q \le t \le t_{q+1}\}}}{e^{\beta_j Z}}\right]$$

$$\left(\frac{\partial AR_{jq}}{\partial P_{jq}}\right) = \frac{1}{N \cdot P_{jq}^2}\left(\int \frac{I_{\{S=j\}}\ I_{\{t_q \le t \le t_{q+1}\}}}{e^{\beta_j Z}}\ d\left\{F_0\right\}\right) = \frac{1}{N \cdot P_{jq}^2} E\left[\frac{I_{\{S=j\}}\ I_{\{t_q \le t \le t_{q+1}\}}}{e^{\beta_j Z}}\right].$$

The influence function ψ_{bji} for $\hat{\beta}_j$ should be obtained as previously discussed. Deriving the influence function for $\hat{P}_{jq}$ is straightforward in that the estimator is linear, needing only to be centered to yield $\psi_{pjqi} = I_{\{S_i=j\}} I_{\{t_q \le t_i \le t_{q+1}\}} - E\left[I_{\{S=j\}} I_{\{t_q \le t \le t_{q+1}\}}\right]$.

The last component needed for the influence function of $\hat{AR}_{jq}$ is the influence function for the empirical distribution, $\phi'_{F_0}\left(F_n - F_0\right) = \frac{1}{N}\sum_{i=1}^{N}\psi_{eji}$. We solve for ψ_{eji} using the same approach as employed in Appendix E, expressing $\hat{AR}_{jq}$ and AR_{jq} as a function ϕ of the empirical distribution F_n and the true distribution F_0 respectively, and taking the derivative in the direction of the true distribution through a scalar

parameter t. In this case, the calculation is

$$\psi_{eji} = \phi'_{F_0}\left(\delta_{S_i,T_i,Z_i}(S,T,Z) - F_0\right) = \frac{d}{dt}|_{t=0}\ \phi\left((1-t)\ F_0 + t\ \delta_{S_i,T_i,Z_i}(S,T,Z)\right)$$

$$\psi_{eji} = \frac{d}{dt}|_{t=0}\ \left[1 - \frac{1}{N\cdot P_{jq}}\left(\int \frac{I_{\{S=j\}}\ I_{\{t_q\le t\le t_{q+1}\}}}{e^{\beta_j Z}}\ d\left\{(1-t)\ F_0 + t\ \delta_{S_i,T_i,Z_i}(S,T,Z)\right\}\right)\right]$$

$$\psi_{eji} = \frac{d}{dt}|_{t=0}\ \left[1 - \frac{1-t}{N\cdot P_{jq}}\left(\int \frac{I_{\{S=j\}}\ I_{\{t_q\le t\le t_{q+1}\}}}{e^{\beta_j Z}}\ d\left\{F_0\right\}\right)\right.$$

$$\left. -\frac{t}{N\cdot P_{jq}}\left(\int \frac{I_{\{S=j\}}\ I_{\{t_q\le t\le t_{q+1}\}}}{e^{\beta_j Z}}\ d\left\{\delta_{S_i,T_i,Z_i}(S,T,Z)\right\}\right)\right]$$

$$\psi_{eji} = \frac{1}{N\cdot P_{jq}}\left(E\left[\frac{I_{\{S=j\}}\ I_{\{t_q\le t\le t_{q+1}\}}}{e^{\beta_j Z}}\ \right] - \frac{I_{\{S_i=j\}}\ I_{\{t_q\le t_i\le t_{q+1}\}}}{e^{\beta_j Z_i}}\ \right).$$

Putting these components together, and replacing true values with estimates and expectations with empirical means, we can compute the influence function for $\hat{AR}_{jq}$

$$\psi_{ajqi} = \left(\frac{\partial AR_{jq}}{\partial \beta_j}\right)\psi_{bji} + \left(\frac{\partial AR_{jq}}{\partial P_{jq}}\right)\psi_{pjqi} + \psi_{eji} + o_p(1).$$

Chapter 3

Building Calibrated Risk Models by Leveraging Information from Published Models

This chapter contains material in preparation to be published in collaboration with Raymond Carroll and Nilanjan Chatterjee.

3.1 Introduction

In the previous chapter, we presented methodology for building an absolute risk model from scratch by integrating information from multiple data sources. However, in many cases a published risk model based on known risk factors may already exist in the literature. If that is the case, the main reason for building a new absolute risk model would be to update the model to include newly identified risk factors, such as lifestyle factors or biomarkers, along with the existing ones.

In this chapter, we investigate methodology for updating risk models with new

risk factor information while incorporating information from existing models as much as possible. Specifically, we consider a regression calibration estimator that has traditionally been used to increase the efficiency of estimation by calibrating to external data that comes from the same underlying population as the sample data. We seek to understand whether this estimator is also useful for calibration in contexts where the two populations are not the same.

The main question we address is whether the regression calibration estimator produces meaningful results when the sample and the external data are representative of different populations, and under what conditions. First, we describe the statistical formulation of the problem along with a few motivating examples and set up basic notations. We then review the regression calibration approach and go on to show analytically that applying the regression calibration estimator when the two populations are different produces an estimate of the external population parameters under certain conditions. In particular, we identify a key mapping that implicitly relates parameters of interest through the population distribution. We show that if this mapping is common for the two populations, then the regression estimator is unbiased for the external population parameters up to a Taylor's approximation. In addition, we provide a variance estimator that is appropriate for this setting. We also conduct extensive simulations to assess the estimator's bias and variance in a variety of settings, numerically corroborating our analytic results. Finally, we discuss exciting areas of future research identified by our work with the calibration estimator.

3.2 Methods

Models and Notations

Let Y be an outcome of interest and X be a set of covariates upon which a published model for the predictive distribution $g(y|x)$ has been built. In general, we will assume that we have only access to the model, but not necessarily to the individual level data from the "external study" upon which the original model was built. Let Z be a set of new covariates based on which the model needs to be updated. We assume that data on Y, X and Z are available to us from an "internal study" for building such a model. We are interested in the case where our internal study sample, which is representative of some underlying population P^I, may differ in some respects from the underlying population that we want to model, P^E, due to characteristics of the study design or sampled population. We envision a situation where the external, published study is representative of the population of interest, P^E, and can thus help produce results that are more generalizable to that population. To make this less abstract, we present two specific examples.

Example 1.

It is often of interest to develop a logistic model for disease risk prediction of the form: logit $\mathrm{P}(D = 1|Z, X) = \beta_0 + \beta_X X + \beta_Z Z + \beta_{XZ} XZ$, where X are established environmental and lifestyle related risk factors and Z is a set of new biomarkers, such as genetic susceptibility markers.

In a logistic model, the intercept parameter β_0 captures information related to the disease rate in the population. It is well known that under case-control sampling the estimated intercept parameter is not unbiased for β_0, but instead for $\beta_0^* = \beta_0 + \log(\pi_1/\pi_0)$, where π_1 and π_0 denote the sampling probability for cases and controls

from the underlying population. To make adequate risk predictions, the estimated intercept parameter must be representative of the disease rate in the underlying population. To obtain a corrected estimate that reflects the true underlying β_0, one can benefit by leveraging external information in the estimation of $\hat{\beta}_0$.

Initial biomarker data are often collected in case-control studies, where differential participation of cases and controls by factors related to lifestyle or behavioral factors can lead to selection bias in the associated risk parameter estimates, β_X. However, if the selection of participants does not depend on the biomarker Z conditional on lifestyle factors X, one can still obtain unbiased estimates for the risk parameters β_Z and β_{XZ} from a case-control study. If that is the case, it may be desirable to estimate the parameters β_Z and β_{XZ} from the case-control study, but utilize information from external models to assist in estimation of β_X, which is susceptible to selection bias in the internal case-control data.

Example 2.

For late-onset chronic diseases, risk models typically require specification of an age-window over which prediction is desired. In this setting it is natural to consider models that incorporate time-to-event as an outcome. To that end, it is common to develop Cox proportional hazard models of the form $\lambda(t|X) = \lambda_0(t)\exp(\beta^T X)$, where $\lambda(\cdot|X)$ denotes the instantaneous hazard function given risk factor information X, $\lambda_0(t)$ denotes the baseline hazard function and β denotes the vector of hazard-ratio parameters. As we described in Section 2.2, hazard ratio parameters β can be estimated from cohort studies, case-cohort studies or even case-control studies that use appropriate incident density sampling design by applying Cox partial-likelihood or conditional logistic regression methods. These hazard ratios β are generally thought to be transportable across populations even though the distributions of risk factors

may differ.

Estimates of $\lambda_0(t)$ can obtained from such studies using Breslow's method or variations of it that can take into accounting the sampling of cases and controls within a cohort. However, if the internal study data is not representative of the population of interest then the resulting baseline hazard rates $\lambda_0(t)$ will also not be representative. For example, Nurses Health Study is a widely used cohort study that has been used for etiologic investigation of many disease outcomes (Colditz et al., 1997). Although this study has proven to be a tremendous resource for understanding relative risks associated with various etiologic causes of various outcomes, any estimate of baseline risks $\lambda_0(t)$ from this cohort would probably not be representative of the general US population because the cohort of participating nurses is likely healthier. If the goal is to build a risk model for the US population using the Nurses Health Study cohort, one would benefit from incorporating external "models" (perhaps just simple marginal disease rates) when estimating $\lambda_0(t)$ part of the model.

We have talked about the fact that national disease registries, such as the SEER registry for cancer, are excellent sources of external data for building risk models in this context. Registry data can be used to obtain estimates of nationally representative age-specific hazard rates for many diseases. Gail et al. (1989) pioneered a method for utilizing this type of external data to estimate the baseline hazard $\lambda_0(t)$ such that the corresponding marginal hazard function (i.e. the hazard function averaged over the covariate distribution) is calibrated to population hazard estimates provided by the registry. In Chapter 2 we extended those ideas to the setting of subtype-specific risk models. In fact, both of these approaches, which build absolute risk models that are in some sense "calibrated" to registry data, are special cases of the general problem we have described.

Statistical Framework

Our goal is to develop a parametric model of the form $f_{\beta_x^E,\beta_z^E,\beta_{zx}^E}(Y|X,Z)$ for the population of interest P^E. We explore doing this by using a sample of population P^I and calibrating to external information on P^E. β_x^E denotes a set of parameters associated with the original risk factors X, β_z^E denotes the main effect parameters associated with new risk factors Z, and β_{xz}^E denotes the possible interactions of Z with the existing risk factors X. A model of the same form for population P^I is parametrized by β_x^I, β_z^I, and β_{xz}^I. The underlying parameters β_x^E and β_x^I are not necessarily the same for the two populations. We do, however, assume the parameters $[\beta_z^E, \beta_{xz}^E] = [\beta_z^I, \beta_{xz}^I]$, so that β_z, β_{xz} are transportable between the populations. This assumption is needed in order to build the model as the external study provides no information on the new covariates Z.

3.2.1 Background: Links with Survey Methodology

Thus far we have spoken generally of calibrating to the information in existing models and external datasets when developing an updated risk model. The problem of how to perform such a calibration has strong links to a well-developed statistical area, namely survey methodology, that we can draw upon. There is a wide literature about how to use external data sources to improve inference for parameter estimates from sample surveys. For a simple introduction to the methodology, consider the problem of estimating the population mean μ_W from a survey that collects random variables W and auxiliary variables V when the population mean μ_V is available from a census. In this setting, commonly used calibration estimators take the form

$$\hat{\mu}_W^{Cal} = \overline{W} + c \times (\mu_V - \overline{V}), \tag{3.1}$$

where $\overline{W}$ and $\overline{V}$ are sample means for W and V, possibly taking into account sample-weights if non-random sampling is used, and c is a constant factor. Since $(\overline{V} - \mu_V)$ is an unbiased estimator for 0 and $\overline{W}$ is unbiased for μ_W, for any constant c, $\hat{\mu}_W^{Cal}$ is an unbiased estimator of the desired quantity, μ_W. The optimal estimator, which minimizes the variance in this class, is given by choosing $c_{opt} = cov(\overline{V}, \overline{W})/var(\overline{V})$. Calibration estimators increase the efficiency of parameter estimation for μ_W by leveraging information on the relationship between W and V in the data and knowledge of μ_V.

The same basic idea can be applied to the problem of calibrating a new model to an existing one. In the model framework described previously, suppose the reduced model $g(y|x)$ is specified in terms of a set of parameters θ. Let $\hat{\theta}^E$ and $\hat{\theta}^I$ denote the maximum likelihood estimates from fitting a reduced model to the external and internal study respectively; analogously, let $\hat{\beta}^I$ denote the maximum likelihood estimate from fitting the full model $f_\beta(y|x,z)$ to the internal study. Denote the corresponding score functions as $U(Y|X;\theta) = \partial \log g(y|x;\theta)/\partial\theta$ and $S(Y|X,Z;\beta) = \partial \log f(y|x,z;\beta)/\partial\beta$ for the reduced and full models respectively.

Analogously to the estimator defined in (3.1), a calibration estimator can be defined as Chen and Chen (2000):

$$\hat{\beta}_{cal} = \hat{\beta}^I + D_1^{-1} C_{12} C_{22}^{-1} D_2 (\hat{\theta}^E - \hat{\theta}^I), \tag{3.2}$$

where

$$D_1 = -E_{S_I}\left[\frac{\partial S(Y|X,Z;\beta)}{\partial \beta^T}\right], \;\; D_2 = -E_{S_I}\left[\frac{\partial U(Y|X;\theta)}{\partial \theta^T}\right],$$

$$C_{22} = E_{S_I}\left[U(Y|X;\theta)U^T(Y|X;\theta)\right], \;\; C_{12} = E_{S_I}\left[S(Y|X,Z;\beta)U^T(Y|X;\theta)\right],$$

and E_{S_I} denotes the sample expectation based on the internal study. It is easy to see

that the maximum likelihood estimators $\hat{\beta}^I$ and $\hat{\theta}_I$ can be asymptotically represented as sample means of the form

$$\hat{\beta}^I = D_1^{-1} E_{S_I} S(Y|X, Z; \beta), \quad \text{and} \quad \hat{\theta}_I = D_2^{-1} E_{S_I} U(Y|X; \theta).$$

Thus, the estimator proposed by Chen and Chen is essentially same as the regression calibration estimator of the form (3.1), but substituting $W = D_1^{-1} S(Y|X, Z; \beta)$ and $V = D_2^{-1} U(Y|X; \theta)$. When the internal and external populations are identical, it is evident that the estimator (3.1) is asymptotically unbiased because $\widehat{\beta}^I$ is a consistent estimator of $\beta^E = \beta^I$ and $(\widehat{\theta}^E - \widehat{\theta}^I)$ is a consistent estimator of zero. However, we are interested in the more general problem of calibration when the internal and external populations may differ in some respects, and it is not clear what the asymptotic limit of β_{cal} is when the underlying populations are different.

3.3 Characterizing Bias for the Calibration Estimator

In the following, we show that under certain conditions the asymptotic limit for the calibration estimator provides a first-order Taylor's approximation for β^E.

We assume that $f(Y|X, Z, \beta)$ specifies a correct model for the conditional distribution $\mathrm{P}(Y|X, Z)$ for both the internal and external study populations, which have parameters β^I and β^E respectively; we will refer to this assumption as (A1). We also assume that $c_{22} = E_{P_I}\left[U(Y|X; \theta)U^T(Y|X; \theta)\right]$ is invertible, where E_{P_I} denotes the population expectation based on the internal study; this is assumption (A2).

Let $q(\beta) = \theta$ define a mapping between the limiting values of the maximum-likelihood estimates $\widehat{\theta}$ and $\widehat{\beta}$. In Appendix A, we show that such a mapping can be

implicitly defined by the equation

$$\int_{Y,X,Z} U(Y|X;\theta)\ f(Y|Z,X;\beta)\ P(Z|X)\ P(X)\ dY\,dX\,dZ = 0. \tag{3.3}$$

The mapping can be common for the internal and external populations, even when the underlying limiting parameter values are not necessarily the same for the two populations, i.e., $q(\beta^I) = \theta^I$ and $q(\beta^E) = \theta_E$ but possibly $\beta^I \neq \beta^E$ or $\theta^I \neq \theta^E$. This mapping plays a key role in understanding whether the calibration estimator will estimate meaningful parameters in a given context.

Proposition. *Assume (A1), (A2). If the mapping $\theta = q(\beta)$ defined by equation (3.3) is the same between populations P^E and P^I, then the calibration estimator $\hat{\beta}_{\text{cal}}$, as defined in equation (3.11), provides an unbiased estimator for β_E, the parameters of the external population, up to a one-step Taylor's approximation.*

Sketch of the Proof. Beginning with equation (3.11) and incorporating the mapping given by equation (3.3), we can write the asymptotic limit of $\widehat{\beta}_{\text{cal}}$ as

$$\beta_{\text{cal}} = \beta^I + d_1^{-1} c_{12} c_{22}^{-1} d_2 \left\{ q(\beta^E) - q(\beta^I) \right\}, \tag{3.4}$$

where d_1, d_2, c_{12}, c_{22} are defined the same way as D_1, D_2, C_{12} and C_{22} respectively, but replacing the sample expectation E_{S_I} by the population expectation E_{P_I}.

In order to understand β_{cal} better, we establish some useful relationships. First, we note the first-order Taylor approximation,

$$q(\beta^E) - q(\beta^I) \approx \left\{ \frac{\partial q(\beta)}{\partial \beta^T}|_{\beta=\beta_I} \right\} \left(\beta^E - \beta^I \right). \tag{3.5}$$

In Appendix B, we take the derivative of the implicit mapping and show that

$$\frac{\partial q(\beta^I)}{\partial \beta^T} = d_2^{-1} c_{12}^T. \tag{3.6}$$

By equating two possible expressions for the asymptotic variance of $\hat{\theta}$, namely $AV\left[\widehat{\theta^I}\right]$ and $AV\left[q(\widehat{\beta^I})\right]$, we show in Appendix C that as long as c_{22} is invertible then

$$d_1 = c_{12}c_{22}^{-1}c_{12}^T. \tag{3.7}$$

Incorporating the relationships (3.5), (3.6), and (3.7) sequentially into (3.4), we show that the asymptotic limit of $\widehat{\beta}_{\text{cal}}$ is

$$\begin{aligned}
\beta_{\text{cal}} &= \beta^I + d_1^{-1}c_{12}c_{22}^{-1}d_2\left\{q(\beta^E) - q(\beta^I)\right\} \\
&\approx \beta^I + d_1^{-1}c_{12}c_{22}^{-1}d_2\left\{\frac{\partial q(\beta)}{\partial \beta^T}|_{\beta=\beta_I}\right\}\left(\beta^E - \beta^I\right) \qquad (3.8) \\
&= \beta^I + d_1^{-1}c_{12}c_{22}^{-1}d_2d_2^{-1}c_{12}^T\left(\beta^E - \beta^I\right) \\
&= \beta^I + d_1^{-1}c_{12}c_{22}^{-1}c_{12}^T\left(\beta^E - \beta^I\right) \\
&= \beta^I + \left(c_{12}c_{22}^{-1}c_{12}^T\right)^{-1}c_{12}c_{22}^{-1}c_{12}^T\left(\beta^E - \beta^I\right) \\
&= \beta^I + \left(\beta^E - \beta^I\right) = \beta^E. \quad \square
\end{aligned}$$

Thus, $\widehat{\beta}_{\text{cal}}$ provides a first-order Taylor's approximation for β^E.

The Taylor's approximation incorporated in (3.8) is based on the relationship $q(\beta^E) - q(\beta^I) \approx \left\{\frac{\partial q(\beta)}{\partial \beta^T}|_{\beta=\beta_I}\right\}\left(\beta^E - \beta^I\right)$; however, by the mean value theorem we know that this relationship holds with equality, i.e. $q(\beta^E) - q(\beta^I) = \left\{\frac{\partial q(\beta)}{\partial \beta^T}|_{\beta=\beta^*}\right\}\left(\beta^E - \beta^I\right)$, for some $\beta^* \in [\beta_E, \beta_I]$. When the form of $q(\beta)$ is linear, the derivative $\frac{\partial q(\beta)}{\partial \beta^T}$ is a constant and thus the relationship in (3.8) is exact. Correspondingly, the closer $q(\beta)$ is to being linear, the better the approximation. However, in situations where $q(\beta)$ is non-linear and $\beta^E \neq \beta^I$, the relationship will always be approximate, even asymptotically. In Section 3.5 we examine the performance of this approximation in

a number of realistic scenarios by way of numerical simulation.

Conditions for the Implicit Mapping to be Common

To determine when the mapping is common for the internal and external populations, we examine each element of (3.3) that defines the implicit mapping. The forms of the reduced model and the full model are the same for the internal and external populations, so the $U(Y|X;\theta)$ and $f(Y|Z,X;\beta)$ parts of the mapping are always common. Thus, for the entire mapping to be common, in general one needs $P^I(Z|X) = P^E(Z|X)$ and $P^I(X) = P^E(X)$, which is to say that the joint distribution of X and Z must be the same for the two populations.

In Appendix D we consider the special case where the reduced model $g(Y|X,\theta)$ is also correctly specified, such as when the model is saturated, and show that the mapping can be implicitly defined by

$$\int_{Y,Z} U(Y|X;\theta)\ \ f(Y|Z,X;\beta)\ \ P(Z|X) \quad dY\,dZ = 0 \quad \text{for each value of X.}$$

This mapping does not depend on $P(X)$, so if the reduced model is correctly specified it need only be the case that $P^I(Z|X) = P^E(Z|X)$ for the mapping to be common for the two populations. This underscores that $P^I(Z|X) = P^E(Z|X)$ is the more critical assumption, a finding which we later confirm by simulation.

It is worth noting that at no point, in either the special or the general case, was it necessary to make assumptions about the relationship between $\left[\beta_0^I, \beta_x^I\right]$ and $\left[\beta_0^E, \beta_x^E\right]$ or between θ^I and θ^E for the mapping to be common. It is rather remarkable that these risk parameters and disease rates for the internal and external populations could be quite different and yet the calibration estimator still provides a good approximation of the full model risk parameters for the external population.

3.4 Variance Estimation

Deriving the variance estimator in this context is fairly straightforward. To simplify notation, we express the estimator as

$$\hat{\beta}_{cal} = \hat{\beta}^I + M(\hat{\theta}^E - \hat{\theta}^I) \text{ where } M = D_1^{-1} C_{12} C_{22}^{-1} D_2.$$

The elements of M are Fisher's information matrices and converge more quickly than the maximum likelihood estimators, so here we treat them as constant matrices. Thus, in derving the variance of $\hat{\beta}_{cal}$ we begin with

$$Var\left[\hat{\beta}_{cal}\right] = Var\left[\hat{\beta}^I\right] + Var\left[M(\hat{\theta}^E - \hat{\theta}^I)\right] + 2 \cdot Cov\left[\hat{\beta}^I,\ M(\hat{\theta}^E - \hat{\theta}^I)\right].$$

The internal estimates $\hat{\beta}^I$, $\hat{\theta}^I$ and the external estimates $\hat{\theta}^E$ are estimated on different datasets and thus have zero covariance, resulting in

$$Var\left[\hat{\beta}_{cal}\right] = Var\left[\hat{\beta}^I\right] + M\left(Var\left[\hat{\theta}^E\right] + Var\left[\hat{\theta}^I\right]\right) M^T - 2 \cdot Cov\left[\hat{\beta}^I, \hat{\theta}^I\right] M^T.$$

Letting Σ_E denote the robust variance estimator for the existing model $\hat{\theta}^E$, and inserting robust variance estimates for the remaining components, we obtain

$$Var\left[\hat{\beta}_{cal}\right] = D_1^{-1} C_{11} D_1^{-1} + M\left(\Sigma^E + D_2^{-1} C_{22} D_2^{-1}\right) M^T - 2 \cdot (D_1^{-1} C_{12} D_2^{-1}) M^T.$$

Finally, we insert the original matrices for M to obtain a final expression for the variance of β_{cal}.

$$Var\left[\hat{\beta}_{cal}\right] = D_1^{-1} C_{11} D_1^{-1} + \left(D_1^{-1} C_{12} C_{22}^{-1} D_2\right)\left(\Sigma^E + D_2^{-1} C_{22} D_2^{-1}\right)\left(D_1^{-1} C_{12} C_{22}^{-1} D_2\right)^T$$
$$- 2 \cdot (D_1^{-1} C_{12} D_2^{-1})\left(D_1^{-1} C_{12} C_{22}^{-1} D_2\right)^T.$$

This variance expression takes into account the variance in the external model, denoted Σ^E, as well as the covariance between the reduced and full internal models. In the next section we examine the performance of this variance estimator in simulations and show that it results in confidence intervals with appropriate coverage probabilities.

3.5 Simulations

To evaluate the performance of the calibration estimator, we considered a simple simulation setting with an existing risk factor X and a new risk factor Z, related to a binary outcome Y through a full logistic model parametrized by β, $P(Y|X,Z) = expit(\beta_0 + \beta_x X + \beta_z Z + \beta_{xz} XZ)$.

3.5.1 The Binary Covariate Setting

We first consider a setting with binary X and Z, defining the the joint distribution (Y,X,Z) through the parameters $P_x = P(X=1)$, $P_{z0} = (Z=1|X=0)$, $P_{z1} = P(Z=1|X=1)$, and $\beta = [\beta_0, \beta_x, \beta_z, \beta_{xz}]^T$. We defined the true underlying population of interest, P^E, by parameter values P_x^E, P_{z0}^E, P_{z1}^E and $\beta^E = \left[\beta_0^E, \beta_x^E, \beta_z^E, \beta_{xz}^E\right]^T$. Similarly, the underlying distribution represented by the available sample P^I was defined by P_x^I, P_{z0}^I, P_{z1}^I and β^I.

We examined many settings where P^I differed from P^E with respect to one or more of these various features of the population distribution. We classify the different settings into Scenarios 1 through 4, with simulation parameters given in Table 3.1. In Scenario 1, the populations P^I and P^E are identical. In Scenario 2, all features of the two populations are the same except for the risk parameters, $\left[\beta_0^I, \beta_x^I\right]^T \neq \left[\beta_0^E, \beta_x^E\right]^T$.

Table 3.1: Simulation Parameters for the Binary Covariate Setting,
Defining Scenarios in which the Internal and External Populations Differ with Respect to Various Features of the Population Distribution

		Parameters						
		P_x	P_{z0}	P_{z1}	β_0	e^{β_x}	e^{β_z}	$e^{\beta_{xz}}$
External Population	True Values	0.5	0.7	0.3	-3	0.85	2	1.2
Internal Population	Scenario 1	-	-	-	-	-	-	-
	Scenario 2	-	-	-	-2.5	0.6	-	-
	Scenario 3	0.2	-	-	-2.5	0.6	-	-
		0.35	-	-	-2.5	0.6	-	-
		0.5	-	-	-2.5	0.6	-	-
		0.65	-	-	-2.5	0.6	-	-
		0.8	-	-	-2.5	0.6	-	-
	Scenario 4	0.8	0.85	0.15	-2.5	0.6	-	-
		0.8	0.7	0.3	-2.5	0.6	-	-
		0.8	0.55	0.45	-2.5	0.6	-	-
		0.8	0.4	0.6	-2.5	0.6	-	-
		0.8	0.25	0.75	-2.5	0.6	-	-

"-" indicates that the internal and external population parameters are the same

In Scenario 3, both $\left[\beta_0^I, \beta_x^I\right]^T \neq \left[\beta_0^E, \beta_x^E\right]^T$ and $P_x^I \neq P_x^E$. Finally, in Scenario 4 all simulation parameters differ between P^I and P^E, including the conditional distributions $P^I(Z|X)$ and $P^E(Z|X)$. In all simulation settings, we let the parameters associated with the new covariate Z be transportable between the populations; mathematically, $\left[\beta_z^I, \beta_{xz}^I\right]^T = \left[\beta_z^E, \beta_{xz}^E\right]^T$.

For each simulation setting, we conducted 5000 simulations where we generated a cohort of size 150,000 from P^E and obtained an estimate of $\hat{\theta}^E$ for the reduced model $P(Y|X) = expit(\theta_0 + \theta_x X)$, representing the existing model toward which to calibrate. We then generated a case-control sample with 1000 cases and 1000 controls from P^I, according to the parameters of the given simulation setting. We fit the full model $P(Y|X,Z) = expit(\beta_0 + \beta_x X + \beta_z Z + \beta_{xz} XZ)$ from the sample of P^I using both standard logistic regression and the calibration estimator, calibrated to $\hat{\theta}^E$. In the binary covariate setting we have defined, both the full and reduced models are saturated and hence correctly specified. Thus, the conditions for a common mapping are satisfied in all scenarios except Scenario 4.

Investigating Bias in the Binary Covariate Setting

In Table 3.2, we present the bias and mean squared error for $\hat{\beta}_0$ and $\hat{\beta}_x$, with respect to the true parameters in the population of interest, β_0^E and β_x^E, for both the calibration estimator and the standard logistic regression estimator. In these simulations, estimation of $\hat{\beta}_z$ and $\hat{\beta}_{xz}^I$ is not impacted by calibration, so the basic logistic regression estimates and the calibration estimates are identical and unbiased (not presented). Across all simulation settings, the standard logistic regression estimator is biased for the intercept β_0^E. This is to be expected, as it is well known that logistic regression with case-control sampling yields a biased estimate of the intercept parameter. The calibration estimator is based on a case-control sample as well, but by calibrating to

Table 3.2: Percent Bias (Mean Squared Error) of the Calibration Estimator and the Standard Logistic Regression Estimator for Estimating Log Odds Ratios in the External Population, β^E, in the Binary Covariate Setting for Simulation Parameters Specified in Table 3.1

	β_0^E				β_x^E			
	cal		basic		cal		basic	
Scenario 1	0	(0.01)	-88	(6.96)	-1	(0.02)	0	(0.02)
Scenario 2	0	(0.01)	-92	(7.71)	0	(0.02)	214	(0.14)
Scenario 3	0	(0.01)	-86	(6.71)	5	(0.02)	219	(0.16)
	0	(0.01)	-89	(7.19)	1	(0.02)	213	(0.15)
	0	(0.01)	-92	(7.71)	-1	(0.02)	212	(0.14)
	0	(0.02)	-96	(8.33)	1	(0.02)	216	(0.15)
	0	(0.03)	-100	(9.03)	2	(0.03)	216	(0.16)
Scenario 4	3	(0.07)	-102	(9.50)	-151	(0.12)	214	(0.19)
	0	(0.03)	-100	(9.00)	-3	(0.03)	212	(0.16)
	-3	(0.02)	-98	(8.62)	145	(0.07)	217	(0.15)
	-7	(0.05)	-95	(8.20)	284	(0.23)	215	(0.15)
	-10	(0.10)	-93	(7.84)	421	(0.49)	215	(0.15)

"cal" refers to the calibration estimator

"basic" refers to the standard logistic regression estimator

the existing study, the method corrects the intercept estimate to be consistent with the external population. As we expect, in Table 3.2 we see that the calibration of the intercept parameter provides unbiased estimates of β_0^E in all settings except Scenario 4, where we know the common mapping assumption is violated. In Scenario 1 we see that both estimators are unbiased for β_x^E, which is also expected.

In Scenario 2 we see that when the risk parameters differ, the basic logistic estimate is heavily biased for β_x^E. This is the case because it is unbiased for $\beta_x^I \neq \beta_x^E$. The calibration estimator exhibits no bias for either parameter, supporting our analytic finding that the risk parameters do not need to be the same for two populations in order for calibration to be effective.

In Scenario 3, we see that the calibration estimator is unbiased for β_0^E and approximately unbiased for β_x^E over a range of deviations between P_x^I and P_x^E when the risk parameters differ as well. This is consistent with our analytic finding that when the reduced model is correctly specified, we need not assume that P_x is the same for the common mapping assumption to hold. The calibration estimator does exhibit a small degree of bias for β_x^E. This is consistent with our analytic observation that even asymptotically the calibration correction will still be approximate. However, under the specified difference in the risk parameters, we see that this approximation is very good. The standard logistic regression estimator is biased for both parameters for the same reasons as in Scenarios 1 & 2.

In Scenario 4, we know that the assumption of a common mapping does not hold, so we do not expect the calibration estimator to be unbiased. Indeed, we see that as the magnitudes of deviation between $P^I(Z|X)$ and $P^E(Z|X)$ increase, the calibration estimate of β_x^E becomes seriously biased. However, it is worth noting that the degree of bias in estimation of the intercept parameter is relatively small, even when the $P(Z|X)$ distributions are significantly different.

Confidence Interval Coverage in the Binary Covariate Setting

In Table 3.3 we present the percent bias and coverage probability for the estimates of standard deviation associated with the calibration estimates and the standard logistic regression estimates. The estimates of the standard deviation for the calibration

Table 3.3: Percent Bias (Coverage Probability) for the Estimated Standard Errors of the Calibration Estimator and the Standard Logistic Regression Estimator for the Log Odds Ratios in the External Population, β^E, in the Binary Covariate Setting for Simulation Parameters Specified in Table 3.1

	β_0^E				β_x^E			
	cal		basic		cal		basic	
Scenario 1	0	(0.95)	7	(0.00)	1	(0.95)	0	(0.95)
Scenario 2	0	(0.95)	8	(0.00)	0	(0.95)	1	(0.37)
Scenario 3	-1	(0.95)	11	(0.00)	1	(0.95)	0	(0.51)
	0	(0.95)	10	(0.00)	-1	(0.95)	0	(0.39)
	-1	(0.95)	7	(0.00)	-1	(0.95)	-1	(0.38)
	-2	(0.95)	4	(0.00)	-2	(0.95)	-1	(0.42)
	0	(0.95)	3	(0.00)	1	(0.95)	0	(0.56)
Scenario 4	0	(0.94)	1	(0.00)	0	(0.83)	-1	(0.74)
	-1	(0.95)	2	(0.00)	-1	(0.95)	-1	(0.58)
	0	(0.86)	4	(0.00)	0	(0.58)	0	(0.45)
	2	(0.41)	7	(0.00)	1	(0.04)	1	(0.43)
	0	(0.01)	7	(0.00)	0	(0.00)	0	(0.47)

"cal" refers to the calibration estimator

"basic" refers to the standard logistic regression estimator

estimator are given by the variance calculation derived in Section 3.4. These results show that in settings where the calibration point estimates are unbiased, namely Scenarios 1 through 3, the estimates of standard deviation for the calibration estimator

are unbiased as well, with coverage probabilities right at the appropriate 0.95 level. In Scenario 4 where the calibration estimator showed substantial bias, the coverage probabilities are understandably poor. Similarly, in the scenarios where the standard logistic regression point estimates were biased, which include all except the Scenario 1 estimate of β_x^E, the coverage probabilities for the standard logistic regression estimates are unacceptably low.

3.5.2 The Continuous Covariate Setting

Having thoroughly explored the setting of binary covariates, we now turn our attention to the context of continuous covariates. Specifically, we consider the case where X and Z are multivariate normal. To be consistent with our setup for binary covariates, we define the joint distribution $P(Y, X, Z)$ by $P(X) \sim N(\mu_x, \sigma_x)$, $P(Z|X) \sim N\left(\mu_{z|x}, \sigma_{z|x}\right)$, and β as before. In all settings, we set the parameters of the marginal distribution of Z to be $\mu_z = 0$ and $\sigma_z = 0.4$. Given these parameters, it is well known that the parameters of the conditional distribution Z given X are $\mu_{z|x} = \mu_z + \sigma_{zx}^2 \left(\sigma_x^2\right)^{-1} (X - \mu_x)$ and $\sigma_{z|x} = \sigma_z^2 - \sigma_{zx}^2 \left(\sigma_x^2\right)^{-1} \sigma_{xz}^2$, where σ_{xz}^2 is the covariance between X and Z (Seber and Lee, 2003). After computing the parameters for the conditional distribution of Z given X, we choose to either shift the mean $\mu_{z|x}$ by $\delta_{z|x}$ or scale the standard deviation $\sigma_{z|x}$ by $\gamma_{z|x}$. With this setup, the joint distribution $P(Y, X, Z)$ is fully specified by the simulation parameters μ_x, σ_x, σ_{xz}^2, $\delta_{z|x}$, $\gamma_{z|x}$, and β.

We investigate the performance of the calibration estimator in situations where these features may differ between the true underlying population, P^E, and the population represented by the sample, P^I. Again, we organize the simulations into Scenarios 1 through 4 that correspond with those for the binary covariate simulations, and

present the simulation parameters in Table 3.4. In simulating P^E we did not manipulate the distribution of $P(Z|X)$ at all, i.e. $\delta_{z|x} = 0$ and $\gamma_{z|x} = 1$. In Scenario 1 the

Table 3.4: Simulation Parameters for Continuous the Covariate Setting, Defining Scenarios in which the Internal and External Populations Differ with Respect to Various Features of the Population Distribution

		Parameters								
		μ_x	σ_x	σ^2_{xz}	$\delta_{z\|x}$	$\gamma_{z\|x}$	β_0	e^{β_x}	e^{β_z}	$e^{\beta_{xz}}$
External Population	True Values	0.0	0.5	0.1	0	1	-3	0.85	2	1.2
Internal Population	Scenario 1	-	-	-	-	-	-	-	-	-
	Scenario 2	-	-	-	-	-	-2.5	0.6	-	-
	Scenario 3	0.15	-	-	-	-	-2.5	0.6	-	-
		0.30	-	-	-	-	-2.5	0.6	-	-
		0.45	-	-	-	-	-2.5	0.6	-	-
		-	0.3	-	-	-	-2.5	0.6	-	-
		-	0.8	-	-	-	-2.5	0.6	-	-
	Scenario 4	0.3	-	-	0.15	-	-2.5	0.6	-	-
		0.3	-	-	0.3	-	-2.5	0.6	-	-
		0.3	-	-	-	0.7	-2.5	0.6	-	-
		0.3	-	-	-	1.3	-2.5	0.6	-	-
		0.3	-	0	-	-	-2.5	0.6	-	-
		0.3	-	0.19	-	-	-2.5	0.6	-	-

"-" indicates that the internal and external population parameters are the same

underlying populations P^E and P^I are the same. In Scenario 2 the risk parameters for P^I differ from P^E. In Scenario 3 the distribution $P^I(X)$ differs from $P^E(X)$, first by shifting the mean μ_x and then by altering the standard deviation. In all settings in Scenario 3, the risk parameters for the two populations differ as well. In Scenario 4 the distribution $P^I(Z|X)$ differs from $P^E(Z|X)$, first by shifting the conditional mean by $\delta_{z|x}$, then by scaling the conditional standard deviation by $\gamma_{z|x}$, and lastly by changing the covariance of X and Z through parameter σ^2_{xz}. In addition, in Scenario 4 the risk parameters and the distribution of $P(X)$ differ as well.

Again, in each simulation scenario we performed 5000 simulations, generating a cohort study of size 150,000 from P^E and a case-control study with 1000 cases and 1000 controls from P^I, fitting models of the same form as in the binary case. However, with continuous covariates it is no longer the case that the existing reduced model is correctly specified; thus, the common mapping assumption will not satisfied in cases where either $P(X)$ or $P(Z|X)$ differ between the populations (Scenarios 3 and 4). Accordingly, we do not expect the calibration estimator to be unbiased in those settings.

Investigating Bias in the Continuous Covariate Setting

In Table 3.5, we present the bias and mean squared error for the calibration estimator and the standard logistic regresssion estimator with respect to the true parameters in the population of interest P^E for the simulation settings described. In these simulations, the calibration estimate of $\hat{\beta}_z$ is not impacted by calibration and thus the results are identical to those for the standard logistic regression estimator and unbiased in all settings; however, calibration estimate of $\hat{\beta}_{xz}$ *is* affected by calibration in these settings and will be discussed. As before, the basic logistic regression estimator is significantly biased for the intercept parameter due to case-control sampling.

The results in Scenarios 1 and 2 mirror what we observed for binary covariates. In Scenario 1 where the underlying populations are the same, both methods provide unbiased estimation of all parameters, except for the standard logistic regression estimator which is biased for the intercept as previously discussed. Scenario 2 shows that when the risk parameters differ, the calibration method is unbiased for the intercept. However, the calibration estimator of β_x^E shows some bias due to the fact that the calibration is approximate, even asymptotically. We also see that the magnitude of bias in the continuous covariate setting is greater than the bias observed

Table 3.5: Percent Bias (Mean Squared Error) of the Calibration Estimator and the Standard Logistic Regression Estimator for Estimating Log Odds Ratios in the External Population, β^E, in the Continuous Covariate Setting for Simulation Parameters Specified in Table 3.4

	β_0^E				β_x^E				β_z^E				β_{xz}^E			
	cal		basic		cal		basic		cal		basic		cal		basic	
Scenario 1	0	(0.00)	-98	(8.72)	2	(0.00)	-1	(0.01)	0	(0.02)	0	(0.02)	2	(0.04)	1	(0.04)
Scenario 2	0	(0.00)	-98	(8.71)	-12	(0.00)	215	(0.13)	0	(0.02)	0	(0.02)	-1	(0.05)	-1	(0.04)
Scenario 3	-1	(0.00)	-101	(9.16)	-4	(0.00)	215	(0.13)	0	(0.02)	0	(0.02)	1	(0.05)	2	(0.04)
	-3	(0.01)	-103	(9.62)	4	(0.00)	216	(0.13)	0	(0.02)	0	(0.02)	2	(0.04)	2	(0.04)
	-5	(0.02)	-106	(10.09)	12	(0.00)	215	(0.13)	0	(0.03)	0	(0.03)	2	(0.05)	1	(0.04)
	-1	(0.00)	-99	(8.81)	307	(0.30)	211	(0.19)	0	(0.04)	0	(0.04)	0	(0.09)	1	(0.08)
	0	(0.00)	-97	(8.47)	-121	(0.04)	216	(0.13)	0	(0.02)	0	(0.01)	2	(0.03)	0	(0.02)
Scenario 4	1	(0.00)	-100	(8.94)	23	(0.01)	217	(0.14)	0	(0.02)	0	(0.02)	3	(0.04)	3	(0.04)
	4	(0.02)	-96	(8.29)	39	(0.01)	215	(0.14)	0	(0.02)	0	(0.02)	2	(0.04)	2	(0.04)
	-3	(0.01)	-104	(9.71)	0	(0.01)	216	(0.14)	1	(0.04)	1	(0.04)	1	(0.06)	1	(0.06)
	-2	(0.01)	-103	(9.49)	10	(0.00)	214	(0.13)	0	(0.01)	0	(0.01)	4	(0.03)	4	(0.03)
	0	(0.00)	-103	(9.54)	-169	(0.08)	215	(0.13)	0	(0.02)	0	(0.02)	-1	(0.07)	1	(0.05)
	-6	(0.04)	-104	(9.71)	177	(0.16)	217	(0.21)	1	(0.14)	1	(0.13)	0	(0.03)	0	(0.03)

"cal" refers to the calibration estimator
"basic" refers to the standard logistic regression estimator

in the binary covariate case. The standard logistic regression estimate is unbiased for $\beta_x^I \neq \beta_x^E$, so we observe significant bias with respect to β_x^E.

In Scenario 3, we see that when the mean of the distribution $P^I(X)$ is shifted relative to $P^E(X)$, the calibration estimates of β_0 and β_x are biased. The bias in the intercept β_0 is negligible, while the bias for β_x is more substantial, increasing in magnitude as the size of the shift in mean increases. Changing the distribution of $P(X)$ through the standard deviation, such that $\sigma_x^I \neq \sigma_x^E$, does not bias the calibration estimate of the intercept parameter but results in substantial bias in the calibration estimate of β_x. It appears that differences in the mean affect the calibration estimate of the intercept more than differences in the standard deviation of $P(X)$, whereas the opposite is true for estimation of $\hat{\beta}_x$. The calibration estimator shows a small degree of bias for β_{xz}. Standard logistic regression is unaffected by the distribution of the covariates, and accordingly its performance is the same as in Scenario 2 in all cases, with the observed bias attributable to differences in the risk parameters.

In the first two settings of Scenario 4, we observe that differences in the mean of the conditional distribution $P(Z|X)$ result in a small degree of bias in the calibration estimate of β_0 and somewhat greater bias in the calibration estimate of β_x, with bias increasing as the difference in means, $\delta_{z|x}$, increases. However, in the next two settings, we observe that differences in the standard deviations of $P(Z|X)$ do not result in bias for the calibration estimate of β_0 and only negligible bias for the calibration estimate of β_x. Differences in the covariance of X and Z do not particularly impact the calibration estimator of β_0, but do substantially affect estimation of β_x. The calibration estimates of β_{xz} show a small degree of bias when the conditional means differ. As we noted previously, the performance of standard logistic regression does not depend on the distribution of the covariates, so again it performs similarly as in Scenario 2.

Confidence Interval Coverage in the Continuous Covariate Setting

In Table 3.6 we present the percent bias and coverage probabilities for the standard deviation estimates that arise from the variance calculation derived in Section 3.4 in the continuous covariate simulation scenarios. In general, the true standard deviations

Table 3.6: Percent Bias (Coverage Probability) for the Estimated Standard Errors of the Calibration Estimator and the Standard Logistic Regression Estimator for Log Odds Ratios in the External Population, β^E, in the Continuous Covariate Setting for Simulation Parameters Specified in Table 3.4

	β_0^E				β_x^E				β_z^E				β_{xz}^E			
	cal		basic		cal		basic		cal		basic		cal		basic	
Scenario 1	-3	(0.95)	128	(0.00)	-7	(0.93)	-1	(0.95)	-5	(0.94)	-1	(0.95)	-5	(0.94)	0	(0.95)
Scenario 2	-4	(0.94)	119	(0.00)	-7	(0.92)	0	(0.08)	-4	(0.94)	0	(0.95)	-6	(0.94)	-2	(0.95)
Scenario 3	-11	(0.58)	106	(0.00)	-6	(0.93)	0	(0.09)	-2	(0.95)	1	(0.96)	-6	(0.93)	-2	(0.95)
	-13	(0.11)	65	(0.00)	-6	(0.93)	0	(0.08)	-4	(0.94)	-1	(0.95)	-4	(0.94)	0	(0.95)
	-14	(0.01)	38	(0.00)	-6	(0.93)	1	(0.08)	-5	(0.93)	-1	(0.95)	-6	(0.94)	-2	(0.95)
	-5	(0.91)	83	(0.00)	-2	(0.43)	0	(0.76)	-1	(0.95)	1	(0.95)	-5	(0.94)	0	(0.95)
	-9	(0.92)	149	(0.00)	-8	(0.00)	-1	(0.00)	-4	(0.94)	1	(0.95)	-17	(0.90)	0	(0.95)
Scenario 4	-12	(0.83)	82	(0.00)	-1	(0.92)	0	(0.11)	-4	(0.94)	0	(0.95)	-5	(0.94)	-1	(0.95)
	-5	(0.03)	54	(0.00)	2	(0.90)	0	(0.19)	-4	(0.94)	-1	(0.95)	-5	(0.94)	-1	(0.95)
	-11	(0.17)	49	(0.00)	-5	(0.94)	1	(0.15)	-2	(0.95)	1	(0.95)	-4	(0.94)	0	(0.95)
	-13	(0.15)	75	(0.00)	-6	(0.93)	1	(0.07)	-5	(0.94)	0	(0.95)	-7	(0.94)	-2	(0.94)
	0	(0.87)	107	(0.00)	0	(0.00)	1	(0.03)	-7	(0.94)	0	(0.95)	-13	(0.91)	0	(0.95)
	-1	(0.45)	11	(0.00)	-2	(0.81)	-1	(0.77)	-2	(0.95)	-1	(0.95)	-4	(0.95)	-2	(0.95)

"cal" refers to the calibration estimator
"basic" refers to the standard logistic regression estimator

were quite small, so even a very small amount of bias appears large when presented on the scale of percent bias. A better indicator of performance of the variance estimator is to consider the coverage probabilities, which is what we will focus on.

In Scenario 1, we see that both methods have coverage probabilities hovering around the appropriate 0.95 level. In this case the point estimates are unbiased so we would expect the standard deviation estimates to be unbiased as well; however we do see a bit of bias in the variance estimator for the calibration estimator. This is most likely due to the fact that these sample sizes are not quite large enough for the asymptotics of the robust variance estimates to have kicked in yet. In additional simulations (not shown) we found that this bias shrank as we looked at increasingly larger sample sizes, supporting the notion that this is a small sample bias.

As we observed in Table 3.5, in Scenarios 2 through 4 there is some degree of bias in the point estimates of β_0 and β_x for the calibration estimator. In these settings where the point estimates are biased, we expect the variance estimator to show some bias as well. However, even in these cases the coverage probabilities still seem relatively reasonable, with a few exceptions. In Scenario 2, we see that there is a large percent bias in the standard deviation estimates, especially when the risk relationship is reverse, and yet the coverage probabilities for all parameters are still greater than 90%. In Scenario 3, when the populations have different means for $P(X)$, the calibration coverage probabilities for β_x are still above 90% but the variance estimates of the intercept β_0^E are significantly affected, with unacceptably low coverage. Conversely, when the populations have different standard deviations for $P(X)$, the coverage probabilities for β_0^E are above 90% while the coverage probabilites for β_x^E drop significantly.

In Scenario 4, when the distribution of $P(Z|X)$ differs between the populations, for the most part the coverage probabilities are still very good, despite the bias in the point estimates. However, when the means of the distributions are different, the calibration coverage probabilities for β_0 are unacceptably low; and when the covariance between X and Z is 0 in the sample but non-zero in the true population,

then the coverage for β_x is also quite bad. The calibration coverage probabilities for β_z and β_{xz} remain above 85% for all simulation scenarios, and in most cases hover very close to the 0.95 level.

Other Observations

In the course of conducting these simulations, we noted the importance of using a robust variance estimator for the existing model, Σ_E, in the variance calculation for the calibration estimator. Standard logistic regression software in R does not return the robust variance estimator by default, so it is unlikely that the variance reported for a published model is the robust variance estimate unless specifically stated. In practice it may be necessary to contact the researchers who built the existing model to request the robust variance matrix. For the purposes of calibration, it would be even better if researchers made it standard practice to publish the robust variance estimator from the outset.

Our simulation explored the performance of the calibration estimator in great detail for the setting of building a logistic regression model, calibrated to a published logistic regression model. This application of the calibration estimator addresses the motivating problem we described in Example 1 of 3.2. However, the theoretical results we've shown for the calibration estimator hold generally and can be applied to calibrate other types of models as well. For instance, to address the motivating problem we laid out in Example 2 of 3.2, in Appendix E we show how to use calibration estimator to calibrate the Cox Proportional Hazards model.

3.6 Conclusions and Practical Recommendations

We proposed the use of a calibration estimator as a way of incorporating relevant external information when building a new or updated model with the aim of improving the efficiency and representativeness of the model. The estimator had not previously been studied in the setting where the sample data and the external information are representative of different populations, so we evaluated its performance in that context, both analytically and numerically. We identified a mapping, given by equation 3.3, that plays a key role in whether or not the calibration estimator produces meaningful results when the populations are different. We showed that if the mapping is common for the two populations, then the calibration estimator is asymptotically unbiased for the parameters of the external population, up to a Taylor's approximation.

With further exploration of the mapping, we determined that a critical requirement for the mapping to be common, and thus for the calibration estimator to perform well, is that the conditional distribution of the new risk factors, Z, given the published risk factors, X, be the same in the two populations. For a given application, researchers should carefully consider whether this is the case based on the features of their particular dataset and the source of external information that is available.

We also found that if the external model is not correctly specified, then the performance of the calibration estimator is sensitive to differences in the distribution of the published risk factors, X, and is especially impacted when the degree of variation differs between in the two populations. Thus, we recommend calibrating to saturated models when possible. In general, calibrating to external information in a form that is most saturated will reduce the impact that any potential differences in the distribution of the published risk factors, X, would have on the resulting model.

3.7 Future Work

In the course of evaluating the calibration estimator we identified some weaknesses that have inspired avenues of future work in this area. As we have discussed, we found that the performance of the calibration estimator depends on whether the distribution of $P(Z|X)$ in the internal data is representative of the distribution in the external population of interest, and in cases where the external model is not correctly specified, on $P(X)$ as well. Unfortunately, these distributions contribute to the calibration implicitly in a way that does not make it possible to incorporate better estimates of $P(Z|X)$ and $P(X)$, perhaps from survey data, in situations where it would be necessary. This inspired us to consider ways to overcome that limitation, while still making use of the powerful mapping that relates parameters of the new model to the parameters of the existing model.

In the future, we plan to explore a constrained maximum likelihood approach that makes use of the key mapping given by equation (3.3) as a constraint in the likelihood. The mapping will still involve the distributions $P(Z|X)$ and $P(X)$, but an advantage of this approach is that one would have the option of incorporating estimates of those distributions from other sources, such as survey data, into the constraint.

3.8 Appendix A: The Mapping

Here we will derive the implicit mapping $q(\cdot)$ that relates the parameters of the full model with the parameters of the reduced model. We begin with the identity

$$E\left[\ U(Y|X;\theta)\ \right] = 0, \text{ which holds for every true value } \theta.$$

$$\int_{Y,X} U(Y|X;\theta)\ P(Y,X)\ dYdX = 0$$

$$\int_{Y,X,Z} U(Y|X;\theta)\ P(Y,X,Z)\ dYdXdZ = 0$$

$$\int_{Y,X,Z} U(Y|X;\theta)\ P(Y|Z,X)\ P(Z|X)\ P(X)\ dYdXdZ = 0$$

By incorporating the full model parametrized by β for $P(Y|Z,X)$, we establish an implicit relationship with θ defined by

$$\int_{Y,X,Z} U(Y|X;\theta)\ f(Y|Z,X;\beta)\ P(Z|X)\ P(X)\ dYdXdZ = 0$$

3.9 Appendix B: Deriving $\frac{\partial q(\beta^I)}{\partial \beta^T} = d_2^{-1} c_{12}^T$

Here, we derive the relationship $\frac{\partial q(\beta^I)}{\partial \beta^T} = d_2^{-1} c_{12}^T$. We begin with the implicit mapping derived in Appendix A.

$$0 = \int_{Y,X,Z} U(Y|X;\theta)\ f(Y|Z,X;\beta)\ P(Z|X)\ P(X)\ dYdXdZ$$

$$0 = \int_{Y,X} U(Y|X;\theta) \left[\ \int_Z f(Y|Z,X;\beta)\ P(Z|X)dZ\ \right] P(X)\ dYdX$$

To temporarily simplify the notation, express this as $0 = \int_{Y,X} A \cdot B \cdot P^I(X)$, letting $A = U(Y|X, q(\beta^I))$ and $B = \int_Z f(Y|Z, X; \beta)\ P(Z|X)$. Take the derivative of both sides with respect to β.

$$0 = \int_{Y,X} \left(\frac{\partial A}{\partial \beta^T}\right) \cdot B \cdot P^I(X) + \int_{Y,X} A \cdot \left(\frac{\partial B}{\partial \beta^T}\right) \cdot P^I(X)$$

The necessary derivatives are

$$\frac{\partial A}{\partial \beta^T} = \frac{\partial U(Y|X, q(\beta))}{\partial \beta^T} = \frac{\partial U(Y|X, q(\beta))}{\partial \theta} \cdot \frac{\partial \theta}{\partial \beta^T} = \frac{\partial U(Y|X, \theta)}{\partial \theta} \cdot \frac{\partial q(\beta)}{\partial \beta^T},$$

$$\frac{\partial B}{\partial \beta^T} = \int_Z \frac{\partial f(Y|Z, X; \beta)}{\partial \beta^T}\ P(Z|X) = \int_Z S^T(Y|Z, X; \beta)\ f(Y|Z, X; \beta)\ P(Z|X).$$

Plugging in these expressions yields

$$0 = \int_{Y,X} \left(\frac{\partial U(Y|X, \theta)}{\partial \theta} \cdot \frac{\partial q(\beta)}{\partial \beta^T}\right) \cdot \left[\int_Z f(Y|Z, X; \beta)\ P(Z|X)\right] \cdot P(X)\ +$$

$$\int_{Y,X} U(Y|X, q^T(\beta)) \cdot \left(\int_Z S^T(Y|Z, X; \beta)\ f(Y|Z, X; \beta)\ P(Z|X)\right) \cdot P(X).$$

By recognizing that these integrals represent expectations over the population, we can write the equation as

$$0 = E_P \left[\frac{\partial U(Y|X, \theta)}{\partial \theta}\right] \left(\frac{\partial q(\beta)}{\partial \beta^T}\right) + E_P \left[U(Y|X, q(\beta)) S^T(Y|Z, X; \beta)\right].$$

When considering the mapping in the internal population, these population expectations are denoted by the matrices $-d_2$ and c_{12}^T, yielding

$$0 = (-d_2) \left(\frac{\partial q(\beta^I)}{\partial \beta^T}\right) + c_{12}^T.$$

This implies that $\frac{\partial q(\beta^I)}{\partial \beta^T} = d_2^{-1} c_{12}^T$.

3.10 Appendix C: Deriving $d_1 = c_{12}c_{22}^{-1}c_{12}^T$

Let $AV[\cdot]$ denote the asymptotic variance of a given estimator. The mapping $\widehat{\theta^I} = q(\widehat{\beta^I})$ implies that

$$AV\left[\widehat{\theta^I}\right] = AV\left[q(\widehat{\beta^I})\right].$$

$$\text{By the delta method, } \quad AV\left[\widehat{\theta^I}\right] = \left(\frac{\partial q(\beta^I)}{\partial \beta^T}\right) AV\left[\widehat{\beta^I}\right] \left(\frac{\partial q(\beta^I)^T}{\partial \beta}\right).$$

Recall that the matrices d_1, c_{11}, and d_2, c_{22} are representations of the Fisher's information for β and θ respectively. Thus, we can incorporate robust variances for the parameters as follows, yielding

$$d_2^{-1}c_{22}d_2^{-1} = \left(\frac{\partial q(\beta^I)}{\partial \beta^T}\right) d_1^{-1}c_{11}d_1^{-1} \left(\frac{\partial q(\beta^I)^T}{\partial \beta}\right).$$

Plugging in the relationship we derived in Appendix A, that $\frac{\partial q(\beta^I)}{\partial \beta^T} = d_2^{-1}c_{12}^T$,

$$d_2^{-1}c_{22}d_2^{-1} = d_2^{-1}c_{12}^T d_1^{-1}c_{11}d_1^{-1}c_{12}d_2^{-1}$$

$$c_{22} = c_{12}^T d_1^{-1}c_{11}d_1^{-1}c_{12}.$$

We assume the full model $f(y|x,z;\beta^I)$ is correctly specified, which means $d_1 = c_{11}$, and thus

$$c_{22} = c_{12}^T d_1^{-1}c_{12}. \tag{3.9}$$

Recall that the overall matrix, $c = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} d_1 & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$ since $d_1 = c_{11}$.

Let *det* denote the determinant of a matrix. The following are properties of matrix algebra:

$$det\begin{bmatrix} A & B \\ C & D \end{bmatrix} = det[A]det[D - CA^{-1}B], \text{ and when } D \text{ is invertible}$$

$$det\begin{bmatrix} A & B \\ C & D \end{bmatrix} = det[D]det[A - BD^{-1}C].$$

Assuming that c_{22} is invertible and applying these rules to the second expression of the overall matrix c, we see that

$$det[c] = det[c_{22}] \cdot det[d_1 - c_{12}c_{22}^{-1}c_{12}^T] = det[d_1] \cdot det[c_{22} - c_{12}^T d_1^{-1} c_{12}].$$

However, from (3.9) we know that $det[c_{22} - c_{12}^T d_1^{-1} c_{12}] = 0$, meaning that

$$det[c_{22}] \cdot det[d_1 - c_{12}c_{22}^{-1}c_{12}^T] = 0. \tag{3.10}$$

Since we assume c_{22} is invertible, $det[c_{22}] \neq 0$. Thus, in order for (3.10) to be satisfied it must be the case that $det[d_1 - c_{12}c_{22}^{-1}c_{12}^T] = 0$, which implies that $d_1 = c_{12}c_{22}^{-1}c_{12}^T$.

3.11 Appendix D: Mapping in a Special Case

Here we derive the mapping $q(\cdot)$ in the special case where the reduced model is correctly specified, in addition to the full model. In this case, we begin with the identity

$$\int_Y g(Y|X;\theta) = 1, \text{ which holds for all } \theta \text{ and } X.$$

Taking the derivative of both sides yields

$$\int_Y \frac{\partial g(Y|X;\theta)}{\partial \theta} = \int_Y U(Y|X;\theta) g(Y|X;\theta) = 0.$$

Under the assumption that the reduced model is correctly specified, it is the case that $g(Y|X;\theta) = \int_Z f(Y|X,Z;\beta)P(Z|X)dZ$, and thus that

$$\int_Y U(Y|X;\theta) \left[\int_Z f(Y|X,Z;\beta)P(Z|X)dZ \right] = 0$$

defines the mapping for each value of X.

3.12 Appendix E: The Calibration Estimator Applied to the Cox Proportional Hazards Model

In the simulation section, we implemented the calibration estimator in the public health context presented in Example 1, where the objective was to fit a logistic regression model on case-control data while calibrating to an existing reduced model built from a large cohort. Now we turn our attention to Example 2 and detail how to use the calibration estimator to estimate the baseline hazard function for the Cox proportional hazards model, $\lambda(t|X) = \lambda_0(t)e^{X\beta}$, such that it is marginally calibrated to registry incidence rates.

Suppose we have data $(X_i,\ T_i,\ \delta_i)$ from a prospective cohort on covariates X, event time T and censoring indicator δ on N individuals in the United States, which we want to use to fit the the Cox proportional hazards model. Recall that $\lambda(t|X) = P\left(T = t|T \geq t, X\right)$ and the baseline hazard $\lambda_0(t) = P\left(T = t|T \geq t, X = 0\right)$. The SEER cancer registry monitors the US population, and is an excellent source of

precise, representative estimates of cancer incidence, $\hat{P}\left(T=t|T\geq t\right)=\hat{\lambda}_m(t)$. To conceptualize this problem in the framework we have proposed, we think of the model we want to fit, $\lambda(t|X)$, as the full model, and the marginal model $\lambda_m(t)$ that does not incorporate covariates as the reduced model. Thus, the incidence rates from the SEER registry can be thought of as an existing, reduced model toward which to calibrate.

As in Section 2.3.1, we define the hazard functions by parameters at each observed event time t_q for q=1,...,Q, as follows:

$$\lambda_0(t)=\left\{\begin{array}{ll}\lambda_{0q} & \text{for } t_q=t\\ 0 & \text{else}\end{array}\right\};\ \lambda_m(t)=\left\{\begin{array}{ll}\lambda_{mq} & \text{for } t_q=t\\ 0 & \text{else}\end{array}\right\}$$

$$\lambda_0(t)=\sum_{q=1}^{Q}\lambda_{0q}I\left\{t_{0q}=t\right\};\ \lambda_m(t)=\sum_{q=1}^{Q}\lambda_{mq}I\left\{t_{mq}=t\right\}.$$

The calibration estimator for the baseline hazard estimate for the full model in the external population is

$$\hat{\lambda}_{0q,cal}=\hat{\lambda}_{0q}^{I}+D_1^{-1}C_{12}C_{22}^{-1}D_2(\hat{\lambda}_{mq}^{E}-\hat{\lambda}_{mq}^{I}),$$

where $\hat{\lambda}_{0q}^{I}=\frac{d_q}{\sum_{i\in R_q}exp(X_i\beta)}$ and $\hat{\lambda}_{mq}^{I}=\frac{d_q}{\sum_{i\in R_q}1}=\frac{d_q}{n_q}$ are the maximum likelihood estimators estimated in the sample, with d_q denoting the number of observed events at time t_q and n_q are the number at risk at that time. $\hat{\lambda}_{mq}^{E}$ are the incidence estimates provided by the registry at each time. The expressions for Fisher's information matrices in this example are:

$$D_1=-E_{S_I}\left[\frac{\partial S\left(\lambda_{0q}\right)}{\partial\lambda_{0q}}\right];D_2=-E_{S_I}\left[\frac{\partial S\left(\lambda_{mq}\right)}{\partial\lambda_{mq}}\right],$$
$$C_{12}=E_{S_I}\left[S\left(\lambda_{0q}\right)S\left(\lambda_{mq}\right)'\right];C_{22}=E_{S_I}\left[S\left(\lambda_{mq}\right)S\left(\lambda_{mq}\right)'\right],$$

with

$$S\left(\lambda_{0q}\right) = \frac{\delta \cdot I_{\{t_q=t\}}}{\lambda_{0q}} - exp(X\beta) \cdot I_{\{t_q \leq t\}}, \quad \frac{\partial S\left(\lambda_{0q}\right)}{\partial \lambda_{0q}} = \frac{-\delta \cdot I_{\{t_q=t\}}}{\lambda_{0q}^2},$$

$$S\left(\lambda_{mq}\right) = \frac{\delta \cdot I_{\{t_q=t\}}}{\lambda_{mq}} - I_{\{t_q \leq t\}}, \text{ and } \frac{\partial S\left(\lambda_{mq}\right)}{\partial \lambda_{mq}} = \frac{-\delta \cdot I_{\{t_q=t\}}}{\lambda_{mq}^2}.$$

Note that because we have defined these functions by a parameter at each observed event time, the matrices D_1, D_2, C_{12} and C_{22} all have dimension J by J, where J is the number of observed event times. Similarly, the variance matrix Σ_E for the "existing model" parameters would theoretically also be J by J. However, in the case where the incidence rates come from a cancer registry as large as SEER, it may be reasonable to treat the λ_{mq}^E's as known, and effectively without variation, $\Sigma_E = 0$.

Chapter 4

Data Applications and Results

4.1 An Absolute Risk Model for Breast Cancer

4.1.1 Introduction

Thus far, we have reviewed statistical methods for building absolute risk models and developed some novel methodologies that extend the existing methods to accommodate disease subtypes in settings where data sources may be completely or partially missing the disease characteristics that define subtype. We then discussed a method for building risk models that are calibrated to existing published models or to disease registries. Having discussed the statistical methods at length, in this chapter we will apply the methods in a real data setting to develop absolute risk models for breast cancer.

We will begin by developing a model for overall invasive breast cancer based on known breast cancer risk factors and genetic information from 24 single nucleotide polymorphisms (SNPs) using prospective cohort data. Our goal is to use the absolute risk model to project the distribution of breast cancer risk in ages 30-70 for the US population. This provides an opportunity to contrast the results with those from the

subtype-specific model, which is built based on case-control studies in the BCAC. Additionally, we plan to use the risk model to evaluate the potential impact that population-wide risk factor modification would have on the distribution of risk in the United States. Since many standard risk factors for breast cancer are not modifiable, a woman at high risk based on non-modifiable risk factors can only reduce her risk so much, even by adopting the lowest risk health behaviors. With this in mind, we will investigate the impact of risk factor modification for hormone replacement therapy use, body mass index, alcohol consumption, and smoking behaviors on overall breast cancer risk within strata defined by non-modifiable risk.

4.1.2 Materials and Methods

Study Population

We analyzed data on a total of 17,176 invasive breast cancer cases and 19,860 controls from 8 prospective cohort studies participating in the Breast and Prostate Cancer Cohort Consortium (BPC3), including 6 American cohorts (CPSII, NHS, WHS, PLCO, MEC, and WHI), 1 European cohort (EPIC), and 1 Australian cohort (MCCS) (Hunter et al., 2005; Husing et al., 2012). The observations contributed by each cohort are given in Table 4.1. These cohorts contributed information on known breast cancer risk factors including first degree family history, age at menarche, parity, age at first full term birth, menopausal status, age at menopause, height in cm, body mass index (BMI), alcohol consumption in g/day, smoking status, and hormone replacement therapy (HRT) use. Four binary HRT variables included information on ever use of HRT, ever use of estrogen only HRT, ever use of combined estrogen and progestin HRT, and current use of HRT. Additionally, these cohorts included genetic information on 24 SNPs. We excluded 42 cases and 45 controls with risk factor values

Table 4.1: Sample Sizes by Case-Control Status and Cohort in the BPC3 data

	Cases	Controls
CPS2	2558	3215
EPIC	4154	5166
MCCS	930	765
MEC	521	570
NHS	1782	3148
PLCO	790	982
WHI	5772	5349
WHS	669	665
Total	**17176**	**19860**

that differed by more than 4 standard deviations from average after transforming the risk factors to be approximately normal.

Statistical Methods

Completeness of the breast cancer risk factors varied by cohort and is presented in Supplemental Table 4.1. We imputed missing values for all risk factors sequentially, in order of increasing missingness. We constructed each imputation model conditional on case-control status, outcome age, cohort, and all completed variables that were significantly associated with the risk factor being imputed. These imputation models also included any significant two-way interactions between the variables included in the model. In cases where a cohort had no data on a given variable from which to build the imputation model, such as for some HRT variables, the imputation was performed from the model built on the cohort thought to be the most similar based on related variables. We fit logistic models for the association between case-control status and each variable, adjusted for age and cohort, before and after imputation to verify that none of the estimated effects changed by more than 10%, and found that most differed by less than 2% before and after imputation.

Different cohorts also had different patterns of missing data for each of the 24 SNPs in the dataset. Within each cohort, we imputed missing data for each SNP for which there was data from an imputation model conditional on case-control status, family history, and an interaction between the two. In cases where a cohort had no data on a given SNP, we did not attempt to impute a value for that SNP. We planned to incorporate SNP data in the absolute risk model through a single polygenic risk score (PGRS) equal to the sum of estimated log odds ratios for each SNP multiplied by the observed SNP profile for each individual. For the SNPs entirely missing in a given cohort, we imputed the missing component of the polygenic risk score (rather than each individual SNP), taking family history into account using the methods of Chatterjee et al. (2013).

To explain further, to construct a PGRS for the BPC3 data, we fit a logistic regression model with all 24 SNPs, adjusted for family history, categorical age, and cohort for individuals with complete data on those variables to obtain estimates of the log odds ratio parameters for each SNP: $\hat{\beta}_1$, ..., $\hat{\beta_{24}}$. We then used those estimates to compute the $\text{PGRS}_i = \sum_{j=1}^{24} \hat{\beta}_j \cdot \text{G}_{ij}$ for each person using their own particular SNP profile, G_i. However, if a cohort was fully missing data on a given set of SNPs, the PGRS could be decomposed into a component that could be directly computed from the empirical data (SNPs j) and a component that could not (SNPs k), as

$$\text{PGRS}_i = \sum_j \hat{\beta}_j \cdot \text{G}_{ij} + \sum_k \hat{\beta}_k \cdot \text{G}_{ik}.$$

$$\text{PGRS}_i = \sum_j \hat{\beta}_j \cdot \text{G}_{ij} + \gamma_i \quad \text{with} \quad \gamma_i = \sum_k \hat{\beta}_k \cdot \text{G}_{ik}$$

Rather than attempt to impute values for each missing SNP, indexed by k, we instead imputed the missing component of the PGRS, γ_i, still using the empirical SNP data

from each individual for SNPs that did have data, indexed by j. We know that γ_i is a sum of random variables so we can impute from normal distributions with appropriate mean and variance. Chatterjee et al. (2013) detail how to specify the parameters of the imputation in a way that accounts for case-control status and family history. The method relies solely on estimates of the log odds ratios and allele frequencies for the missing SNPs along with an estimate of the log odds ratio for family history, all of which we obtained from the completed data (Chatterjee et al., 2013). For the purposes of comparison, we created a fully simulated PGRS based on 24 SNPs by imputing the entire PGRS, not just the missing component. Using this method, we also created a simulated PGRS based on 86 SNPs, adding an additional 62 SNPs found to be associated with breast cancer in the published literature, with log odds ratios and allele frequencies estimated in the BCAC data.

Using the imputation methods described, we created 5 imputed datasets for analysis. To build the absolute risk model, we employed the methods pioneered by (Gail et al., 1989), the details of which we reviewed in Chapter 2. We estimated the hazard ratio component of the model using standard logistic regression adjusted for quintiles of outcome age and cohort. We estimated the attributable risk among the cases using the Bruzzi method, by computing one minus the average of the inverse relative risks among the cases (Bruzzi et al., 1985). We combined the estimated attributable risk with marginal hazard rates of overall breast cancer in the SEER registry to obtain baseline hazard rates. Having constructed the absolute risk model, we then used it to predict risk of breast cancer among the controls to obtain an estimate of the distribution of risk for ages 30-70 in the population.

When predicting risk over the age interval 30-70, all women are initially premeopausal (with no HRT use) and become postmenopausal at their recorded age at menopause. In addition to menopausal status, at this time the variables reflecting

HRT use are changed from "never" to the actual HRT use behaviors of the women for whom risk is to be predicted. For women who are premenopausal, the age at menopause and HRT use variables are not defined. For the purposes of projecting risk for these women, we use an age at menopause of 50 (the median among postmenopausal women in the dataset) and we impute HRT use variables from a model fit on the postmenopausal women for whom the HRT variables are known. We constructed the absolute risk model and predicted risk for each of the 5 imputed datasets and averaged the results.

4.1.3 Results and Discussion

To determine the form of the hazard ratio component of the absolute risk model, we performed a number of exploratory analyses. To evaluate whether the effect of each risk factor could be modeled in a linear fashion, we fit generalized additive models relating case-control status to each continuous variable, adjusted for age and cohort (Hastie and Tibshirani, 1990). The models allowed us to look at flexibly modeled, smoothed covariate effects to ascertain whether linear modeling was appropriate. In general we found non-linearity in the effects and thus chose to take a more non-parametric approach, including categorical versions of the continuous risk factors in the model. We evaluated heterogeneity in the effects across cohorts by creating forest plots and testing for differences in effect size. We did not see statistically significant heterogeneity in the effects by study, except for the age at first full term birth (AFFTB) variable. In this case, all studies were qualitatively consistent in showing that greater AFFTB increased risk but differed in the estimated effect size.

We also evaluated whether to include an interaction between the PGRS and any of the risk factors. We tested interactions between the PGRS and each covariate, modeled continuously to reduce the degrees of freedom and increase the power for

detecting interaction. We first tested interactions between the PGRS and each continuous covariate, adjusting for age and cohort, for the middle 90% of the PGRS. We separately tested interaction in the extremes of PGRS (defined by the upper and lower 5th percentiles) separately, by including a binary indicator of "extreme" (relative to the middle) and testing the interaction with each continuous covariate. We also looked at forest plots and performed statistical tests to evaluate heterogeneity in the effect size for each continuous covariate across deciles of the PGRS. Across these evaulations we did not see consistent evidence of interaction between PGRS and any risk factor and thus did not include any interactions with PGRS. Previous studies have found that the effect of BMI on breast cancer risk is strongest among postmenopausal never-HRT users (Lahmann et al., 2004; Morimoto et al., 2002; Lahmann et al., 2003). We observe this in the BPC3 data as well and accordingly we include an interaction between BMI and HRT in the model.

Based on our initial analyses, we formulated a fully-adjusted model with main effects for the categorical covariates and an interaction between deciles of BMI and ever use of HRT, along with deciles of PGRS, and adjustment for quintiles of age and cohort. Supplemental Table 4.2 contains the hazard ratio estimates from this model with the PGRS based on empirical genotype data for 24 SNPs, which are consistent with previous associations for the standard breast cancer risk factors. Table 4.2 compares the hazard ratio estimates for deciles of the PGRS based on empirical genotype data for 24 SNPs with the hazard ratio estimates for deciles of the simulated PGRS for 24 SNPs in the fully-adjusted model. The estimates correspond closely, providing support for our use of the simulated PGRS for 86 SNPs going forward. We evaluated the discriminatory accuracy of a number of models, including those with risk factors alone, PGRS alone, and risk factors and PGRS together for both the PGRS based on empirical genotype data for 24 SNPs and the simulated PGRS for

Table 4.2: Comparison of Estimated Hazard Ratios for Deciles of PGRS, for PGRS from Empirical Genotype Data for 24 SNPs and Simulated PGRS for 24 SNPs, in Fully-Adjusted Models

	Estimated Hazard Ratios	
	Empirical	Simulated
PGRS Decile 1	1	1
PGRS Decile 2	1.19	1.21
PGRS Decile 3	1.33	1.38
PGRS Decile 4	1.43	1.47
PGRS Decile 5	1.58	1.59
PGRS Decile 6	1.66	1.75
PGRS Decile 7	1.78	1.89
PGRS Decile 8	2.05	2.02
PGRS Decile 9	2.26	2.27
PGRS Decile 10	2.80	2.84
Family History	1.41	1.38

86 SNPs. Figure 4.1 shows the Receiver Operating Characteristic (ROC) curves and the area under the ROC curve (AUC) for these models (Hanley, 1989). We find that the best discriminatory accuracy is provided by a model with risk factors and the simulated PGRS for 86 SNPs, with AUC=0.654.

We predicted absolute risks among controls for ages 30-70 from two fully-adjusted models, one with the PGRS based on empirical genotype data for 24 SNPs and the other with the simulated PGRS for 86 SNPs. Table 4.3 presents average predicted risk within deciles of risk for these two fully-adjusted models. These results show increased risk stratification for the 86 SNP model, particularly for those at highest risk. Figure 4.2 shows the distribution of predicted absolute risks for ages 30-70 for the fully-adjusted model with simulated PGRS based on 86 SNPs, along with the percentage of the population projected to exceed referent risk thresholds.

Ideally, individuals at high risk of disease could seek to reduce their risk by modifying their behaviors. Unfortunately, most known breast cancer risk factors such as

Figure 4.1: ROC Plot for Risk Models in BPC3

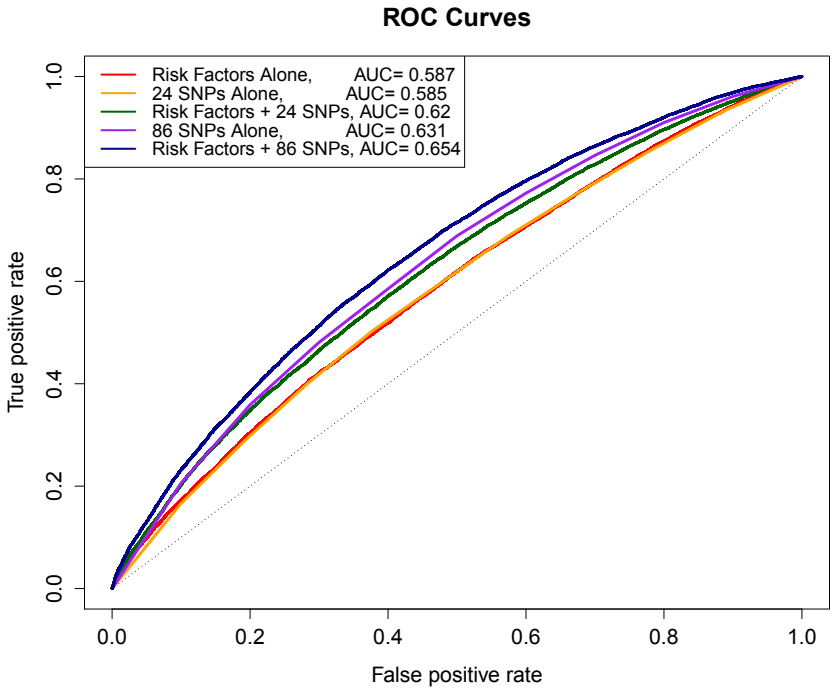
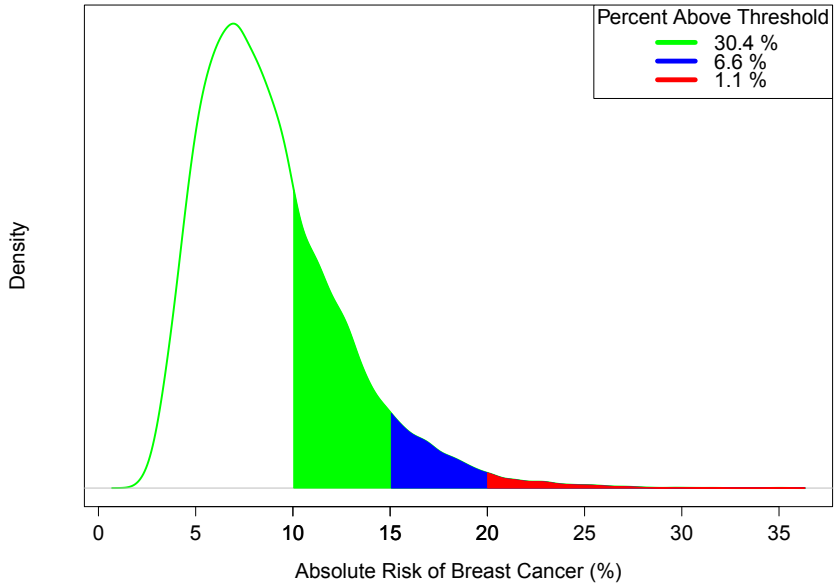


Table 4.3: Average Absolute Risk of Breast Cancer in Ages 30-70 from Two Fully-Adjusted Models, with the PGRS from Empirical Genotype Data for 24 SNPs and the Simulated PGRS for 86 SNPs, by Risk Decile

	Among Controls	
	24 SNPS	86 SNPS
Risk Decile 1	4.19	3.15
Risk Decile 2	5.36	4.35
Risk Decile 3	6.17	5.29
Risk Decile 4	6.93	6.20
Risk Decile 5	7.69	7.15
Risk Decile 6	8.53	8.21
Risk Decile 7	9.52	9.45
Risk Decile 8	10.77	11.13
Risk Decile 9	12.60	13.67
Risk Decile 10	16.98	19.58
Overall	8.87	8.82

Figure 4.2: Distribution of Absolute Risk of Breast Cancer in Ages 30-70 from Fully-Adjusted Model with Simulated PGRS for 86 SNPs



family history, reproductive factors, and genetic factors are not modifiable. In fact, in our absolute risk model the only modifiable risk factors are alcohol consumption, HRT use, BMI, and smoking status. We evaluated how modifying these risk factors to their lowest risk levels, individually and simultaneously, impacts the projected risk distribution for ages 30-70 using the fully-adjusted model with 86 SNPs. In particular, we computed two quantities to quantify the risk reduction associated with modifying risk factors: percent total cancer (PTC) reduction and percent preventable cancer (PPC) reduction.

We defined PTC as the difference between the probability of disease in the population and the probability of disease for those with lowest risk modifiable factors, M_0, as a proportion of the overall probability of disease; mathematically, $PTC = [P(D) - P(D|M_0)]/P(D)$. We then partitioned the PTC by categories defined by

non-modifiable risk group, denoted by G, to be

$$PTC = \left[\sum_{G} \left[P(D|G) - P(D|G, M_0)\right] P(G)\right] / P(D).$$

Thus, the percent total cancer reduction that would occur by reducing all modifiable risk factors to the lowest risk levels in a particular non-modifiable risk category, G', is given by

$$PTC(G') = \frac{\left[P(D|G') - P(D|G', M_0)\right] P(G')}{P(D)}.$$

We defined percent preventable cancer similarly, but with $P(D) - P(D|M_0)$ in the denominator in order to measure the risk reduction as a proportion of the possible reduction that could be achieved if the entire population reduced modifiable risk factors to the lowest risk levels.

$$PPC(G') = \frac{\left[P(D|G') - P(D|G', M_0)\right] P(G')}{P(D) - P(D|M_0)}.$$

We computed $P(D|G)$ within strata defined by the quintiles of non-modifiable risk by using the absolute risk model to predict risk for the covariate profiles of controls in the dataset. To compute $P(D|G', M_0)$ we predicted risk for the same covariate profiles but set the modifiable risk factors to their lowest risk levels. Using these quantities, we computed the proportion of all breast cancers that would be prevented by modification of the risk factor (PTC) and the percent that would be prevented if one were to target risk factor modification to those in a given quintile of non-modifiable risk (PPC).

Table 4.4 presents these results. We see that targeting all risk factors simultaneously is projected to result in a 32% reduction of breast cancer in the population, while targeting HRT-use alone would only result in a roughly 17% reduction overall. We also see that by targeting risk factor modification efforts to those in the upper 20th percentile of non-modifiable risk, roughly 35% of the preventable breast cancers

Table 4.4: Percent Preventable and Percent Total Breast Cancer Reduction for Ages 30-70 from Modification of Risk Factors by Non-Modifiable Risk Group, Based on Fully-Adjusted Model with Simulated PGRS for 86 SNPs

Non-Modifiable	Alcohol		HRT		BMI		Smoking		Simultaneous	
Risk	%P	%T	%P	%T	%P	%T	%P	%T	%P	%T
Quintile 1	9.3	0.71	9.1	1.54	10.2	0.87	9.6	0.38	9.6	3.10
Quintile 2	14.0	1.07	13.8	2.34	14.5	1.23	14.3	0.57	14.2	4.58
Quintile 3	18.0	1.38	18.1	3.06	18.4	1.57	18.2	0.73	18.2	5.89
Quintile 4	23.4	1.79	23.4	3.94	23.0	1.96	23.4	0.93	23.3	7.52
Quintile 5	35.4	2.71	35.5	6.00	34.0	2.90	34.5	1.38	34.8	11.25
Overall	–	7.66	–	16.89	–	8.53	–	3.99	–	32.33

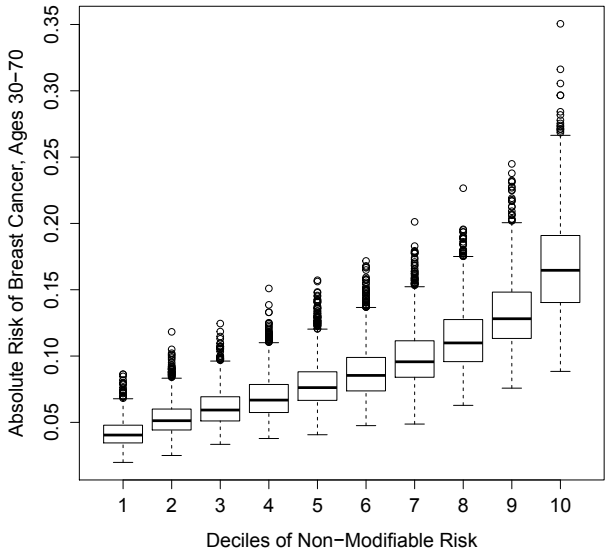
%P refers to percent preventable breast cancer reduction

%T refers to percent total breast cancer reduction

might be prevented. This indicates that targeted prevention strategies may yield a greater breast cancer reduction for the same resources invested. However, this also indicates that a considerable proportion of preventable breast cancer falls outside the highest non-modifiable risk group so there is reason to pursue broader prevention strategies as well.

Within categories defined by non-modifiable risk there is significant variability in overall breast cancer risk, depending on the specific risk factors, both modifiable and non-modifiable, of each woman in that group. Figure 4.3 presents the distribution of breast cancer risk in ages 30-70 by non-modifiable risk group based on results from the fully-adjusted model with simulated PGRS for 86 SNPs. It is clear that the higher risk non-modifiable groups have a greater degree of stratification in breast cancer risk than the lower risk groups. This as an important, though perhaps under-appreciated, consequence of modeling risk with a logistic regression model, which is formulated such that risk factors act multiplicatively on risk. A model with no interaction on the relative risk scale does result in an interaction on the scale of risk differences, which is what we consider when working with absolute risk. Thus, even though there is

Figure 4.3: Distribution of Absolute Risk of Breast Cancer in Ages 30-70 by Non-Modifiable Risk Group, Based on Fully-Adjusted Model with Simulated PGRS for 86 SNPs



no interaction between the modifiable factors and the non-modifiable factors in our logistic model, we see an interaction on absolute risk scale.

This has important ramifications for disease prevention efforts. Though the relative risk of modifying a given risk factor may be the same for two women, their risk differences may be drastically different depending on their other risk factors, with one receiving a substantially greater benefit. This presents a compelling reason for considering disease prevention in public health through the lens of absolute risk.

4.1.4 Conclusions and Future Work

We built two absolute risk models for invasive breast cancer from prospective cohort data in the BPC3 data that incorporated known breast cancer risk factors and polygenic risk scores based on 24 and 86 SNPs respectively. Having developed these absolute risk models, we projected the distribution of breast cancer risk in the BPC3

controls for ages 30-70 and found that approximately 1.1% had a breast cancer risk exceeding 20%. It is possible that the covariate profiles in BPC3 represent women with generally lower risk health behaviors, in which case the projected risk distribution would underestimate the distribution of risk in the general US population. In future work, we plan to obtain nationally representative covariate distributions from the National Health and Nutrition Examination survey (NHANES) to use for projecting risk, and compare the resulting risk distribution to that projected from the BPC3 data.

We used the absolute risk model to investigate the impact of population-wide modification of alcohol consumption, smoking behavior, body mass index, and use of hormone replacement therapy, alone and simultaneously, on breast cancer risk. We evaluated the reduction in breast cancer risk that could be achieved by targeting risk factor modification efforts to women at high risk based on non-modifiable risk factors, such as genetics or defined reproductive risk factors. While we found that targeting risk reduction efforts to the highest risk group based on non-modifiable factors would result in greater breast cancer reduction than intervening in the population at random, we found that much of the preventable breast cancer risk was outside of that high risk group so broader intervention efforts are needed as well.

In the future, we plan to work with collaborators to develop an absolute risk model for invasive breast cancer in the UK. The model will incorporate the hazard ratios presented here, fit on the BPC3 data. However, for the UK absolute risk model we will base the estimates of attributable risk on a risk factor distribution that is representative of women in the UK. Additionally, we will calibrate the model to breast cancer rates from the UK. We plan to validate the model in data from the UK Breakthrough Generations Study.

In the next section, we develop subtype-specific absolute risk models for subtypes

of breast cancer defined by estrogen receptor status in the US population using data from the BCAC consortium of case-control studies. We can obtain estimates of breast cancer risk from the sum of the subtype-specific risks provided by the models. In the future, we also plan to build an absolute risk model for overall breast cancer directly from the BCAC data. We will compare the projected risk distribution of overall breast cancer from BCAC to these BPC3 results. Through this comparison, we will investigate differences that may arise due to building the absolute risk model from a consortium of case-control studies as opposed to a consortium of cohorts. We will also compare the risk stratification produced by the simulated PGRS for 86 SNPs with that from a PGRS based on empirical genotype data on 77 SNPs in the BCAC data.

4.2 Absolute Risk Models for Breast Cancer Subtypes Defined by Estrogen Receptor Status

4.2.1 Introduction

In this section, we make use of the methods developed in Chapter 2 to build absolute risk models for subtypes defined by estrogen receptor (ER) status. ER+ and ER- breast cancer subtypes are quite distinct. In general, ER+ breast cancer has later age of onset compared with ER- breast cancer (Anderson and Matsuno, 2006). ER+ breast cancer is more common, comprising 77% of breast cancers for which ER status is known in SEER (Jatoi et al., 2007). Many risk factors have been identified for ER+ breast cancer, including all reproductive factors included in our previous model for overall breast cancer, but few have been identified for ER- breast cancer (Ma

et al., 2006). ER+ breast cancer is thought to be hormonally-driven, and a number of treatments have been developed, resulting in a generally more favorable prognosis for ER+ compared to ER- breast cancer (Elledge et al., 1994). The most well known of these is tamoxifen, which has also been used as a preventative treatment for ER+ breast cancer (Moyer et al., 2013; Fisher et al., 1998). However, use of tamoxifen has been shown to be associated with adverse events, including stroke, and thus should only be administered as a preventative treatment to women who are expected to derive the most benefit, those at high levels of risk for ER+ breast cancer (Bushnell and Goldstein, 2004). A subtype-specific absolute risk model for ER+ breast cancer can be used to identify women who exceed thresholds of risk for which the benefits of preventative breast cancer treatments outweigh the harms.

In the following sections, we develop subtype-specific risk models for ER+ and ER- breast cancer with the goal of estimating distributions of breast cancer risk for each subtype in the US population, and producing estimates of overall breast cancer risk from the sum of the subtype-specific risks. We then examine and compare the degree of risk stratification produced by these models for ER+, ER-, and overall breast cancer.

4.2.2 Materials and Methods

Study Population

We analyzed data on a total of 36,018 cases of invasive breast cancer and 36,155 controls from 27 case-control studies in the Breast Cancer Association Consortium (BCAC), 10 of which were classified as population based. The BCAC constortium pools information across many large breast cancer studies in order to provide a resource for comprehensive investigation of breast cancer risk factors, in particular

genetic risk factors (Breast Cancer Association Consortium, 2006; Yang et al., 2011b; Milne et al., 2010; Fasching et al., 2012; Nickels et al., 2013). These studies provide information on standard breast cancer risk factors, including first degree family history, parity and age at first full term birth, age at menarche, menopausal status and age at menopause, height, body mass index (BMI), alcohol consumption, smoking behavior, and hormone replacement therapy (HRT) use. Some studies provided more detailed information on HRT, including data on ever use of HRT, current use of any HRT, current use of estrogen-type HRT, and current use of combined estrogen and progestin-type HRT. In addition, all included studies provided genotype data on 77 SNPs previously found to be associated with breast cancer. These SNPs were genotyped using a custom Illumina iSelect genotyping array known as iCOGS, designed as part of the Collaborative Oncological Gene-Environment Study (COGS) to test genetic variants related to breast cancer (Sakoda et al., 2013; Bahcall, 2013). We restricted our analysis dataset to those with iCOGS data, known female sex, and reported European ethnicity. We excluded studies that were classified as family-based or of mixed design.

For cases, additional information was provided on tumor characteristics, indicating whether the tumor was positive, negative or unknown for each of the following markers: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Specifically, the 36,018 cases were distributed as ER+ (65.6%), ER- (15%), unknown ER status (19.4%), with PR+ (48.4%/0, PR- (21.2%), unknown PR status (30.4%); and HER2+ (8.2%/0, HER2- (43.8%), unknown HER2 status (48.0 %). Here, we focus on developing an absolute risk model for breast cancer subtypes defined by ER status, which is the least missing of the tumor characteristics.

Statistical Methods

To be consistent with previous analyses of the BCAC data, we chose to model as categorical the variables family history (no/yes), parity (0/1/2/3/4+), smoking (no/current), and menopausal status and HRT (Premeno, Postmeno Never HRT use, Postmeno Former HRT use, Postmeno Current Non-E-Type HRT use, Postmeno Current Only E-Type HRT use, and Postmeno Current Any EP-Type HRT use. We chose to model as continuous the variables age at menarche divided by 2, age at first full-term birth divided by 5, age at menopause, alcohol consumption divided by 10, and height and bmi divided by 5. All continuous variables were centered by subtracting the median value.

In defining these variables, we encountered two types of missing data. We observed "structural missing" data when studies were entirely missing a given covariate. Any study with fewer than 200 cases or 200 controls with data for a given covariate were treated as structurally missing for that covariate. In other cases, studies had mostly complete data for a given covariate but were "randomly missing" some observations. Due to the large number of studies, we did not attempt to build models to impute missing covariate values. Instead we included missing categories for each covariate, paying careful attention to model the variables with missing categories such that the resulting model parameters were interpretable. Those with missing data on any of the 77 genotyped SNPs comprised less than 4% of the study population and were excluded from the analysis.

To build absolute risk models for ER+ and ER- subtypes, we employed the methods described in Chapter 2. In particular, we estimated the hazard ratio component of the model using the reparametrized multinomial model with the ER+ subtype as referent. A strength of this method is that cases contribute to estimation even when ER status is unknown. Additionally, because ER+ is the referent group, a study is

not required to have controls to contribute to the effect estimation. This makes it possible to include all cases from hospital based studies, that may not have representative controls. We will refer to this method as "full data analysis" (FDA). Table 4.5 gives the sample sizes included in FDA by study, with cases differentiated by ER status.

Alternatively, one could perform a "complete case analysis" (CCA) and estimate the model for each subtype individually using standard logistic regression methods, excluding cases for which the ER status is unknown. To contribute to the CCA, studies must have good controls, which further restricts the analysis to data from population based studies. Table 4.6 gives the sample sizes included in CCA for subtypes defined by ER status. Comparing Table 4.5 and Table 4.6, we see that FDA allows the model to be fit from roughly five times the number of cases compared to CCA.

We estimate subtype-specific baseline hazard rates for the absolute risk models from the marginal SEER incidence rate for overall breast cancer in conjunction with estimates of attributable risk and observed subtype proportions in the data. We estimated attributable risk from cases with complete covariates in 4 time intervals, defined separately for each subtype by the quartiles of outcome age using the Bruzzi formula (Bruzzi et al., 1985). We estimated subtype proportions in 4 time intervals defined by outcome age as well. For ER+ and ER- subtypes, one could instead choose to calibrate to marginal subtype-specific hazard rates directly, which are available in SEER. This is generally a more optimal strategy; however, here we choose to estimate the subtype proportions to provide a real data example of the methods presented in Chapter 2 and to demonstrate what one would do if the subtypes of interest were such that subtype-specific rates were not provided by SEER.

Table 4.5: BCAC Sample Sizes Contributing to Full Data Analysis in Multinomial Model, Overall and by Study

	Controls	ER+	ER-	ER Missing
Total	7480	**23624**	**5613**	**6974**
ABCFS	551	456	261	73
ABCS	0	561	195	397
BBCC	0	456	82	16
BIGGS	0	474	146	173
BSUCH	0	531	152	132
CECILE	999	743	130	27
CGPS	0	1919	357	582
CTS	71	0	68	0
ESTHER	502	303	98	71
GENICA	427	328	119	18
HMBCS	0	35	8	645
KBCP	251	288	89	34
LMBC	0	2069	378	169
MARIE	1778	1279	370	7
MCBCS	0	1271	250	25
MCCS	511	352	119	143
MEC	0	412	87	206
MTLGEBCS	0	421	64	4
NBCS	0	617	199	44
NBHS	118	0	125	0
OBCS	0	403	97	0
ORIGO	0	211	68	56
PBCS	424	519	0	0
pKARMA	0	3588	664	301
SASBAC	1378	663	144	356
SBCS	0	358	104	289
SEARCH	0	5130	1170	2796
SZBCS	0	149	51	103
UKBGS	470	88	18	307

Table 4.6: BCAC Sample Sizes Contributing to Complete Case Analysis in Separate Logistic Models, Overall and by Study

Study	Controls	ER+
Total	7291	**5019**
ABCFS	551	456
CECILE	999	743
ESTHER	502	303
GENICA	427	328
KBCP	251	288
MARIE	1778	1279
MCCS	511	352
PBCS	424	519
SASBAC	1378	663
UKBGS	470	88

Study	Controls	ER-
Total	7056	**1541**
ABCFS	551	261
CECILE	999	130
CTS	71	68
ESTHER	502	98
GENICA	427	119
KBCP	251	89
MARIE	1778	370
MCCS	511	119
NBHS	118	125
SASBAC	1378	144
UKBGS	470	18

4.2.3 Results and Discussion

We modeled the hazard ratios for each subtype from the FDA multinomial regression model with main effects for the covariates, an interaction between continuous BMI and menopausal status and HRT use, and 77 SNPs adjusted for deciles of age, study, and 9 principle components scores to account for population substructure. For purposes of comparison, we fit the same models using CCA in separate logistic regression models. The estimated hazard ratios for the covariates are given in Table 4.7 and the estimated hazard ratios for the SNPs are given in Supplemental Table 4.3. In general, the point estimates for FDA and CCA are consistent with one another for both subtypes, with the exception of the association between parity and ER- breast cancer. As we would expect, FDA is more efficient than CCA, with tighter confidence intervals. Testing statistical significant of the SNP effects in the FDA model finds 47 of the 77 SNPs to be significantly associated with ER+ breast cancer, where with CCA only 44 are significant. Similarly, for ER- breast cancer FDA finds that 23 SNPs are statistically

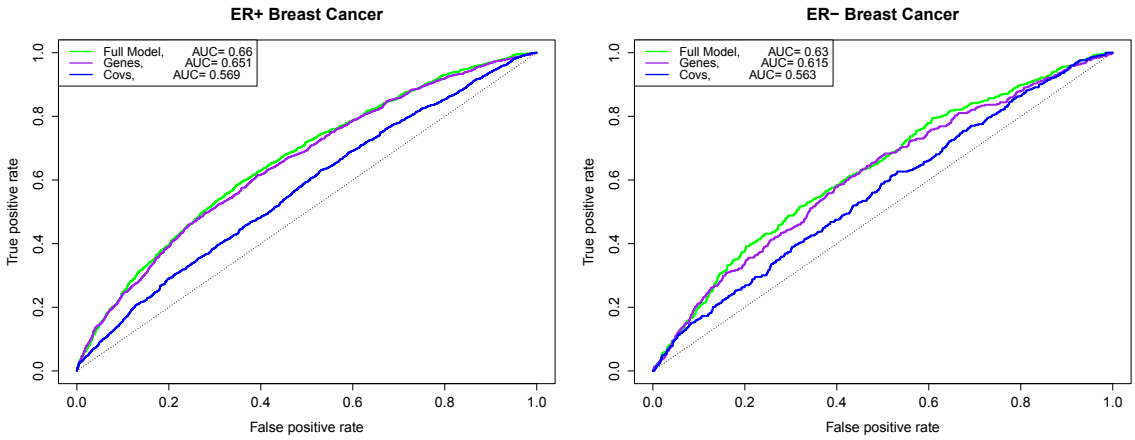
Table 4.7: Covariate Hazard Ratios (HR) and 95% Confidence Intervals (CI) for Fully-Adjusted Models of ER+ and ER- Breast Cancer, with Results from FDA and CCA

	ER+				ER-			
	FDA		CCA		FDA		CCA	
	HR	(CI)	HR	(CI)	HR	(CI)	HR	(CI)
Famiy History = No	1	–	1	–	1	–	1	–
Family History = Yes	1.59	(1.39, 1.81)	1.59	(1.38, 1.82)	1.59	(1.38, 1.84)	1.56	(1.26, 1.94)
Family History = Missing	0.92	(0.72, 1.17)	0.91	(0.7, 1.18)	0.55	(0.43, 0.72)	0.74	(0.46, 1.17)
Parity = 0	1	–	1	–	1	–	1	–
Parity = 1	0.95	(0.84, 1.08)	0.91	(0.79, 1.04)	1.06	(0.91, 1.23)	1.47	(1.18, 1.85)
Parity = 2	0.84	(0.75, 0.94)	0.77	(0.68, 0.88)	0.88	(0.77, 1)	1.21	(0.99, 1.48)
Parity = 3	0.71	(0.63, 0.81)	0.68	(0.59, 0.78)	0.72	(0.62, 0.84)	0.94	(0.74, 1.19)
Parity = 4	0.58	(0.5, 0.67)	0.57	(0.48, 0.68)	0.61	(0.51, 0.73)	0.69	(0.52, 0.92)
Parity = Missing	0.41	(0.21, 0.78)	0.36	(0.17, 0.77)	0.24	(0.12, 0.47)	0.4	(0.11, 1.39)
Age Menarche /2	0.93	(0.89, 0.97)	0.93	(0.88, 0.97)	0.93	(0.88, 0.98)	0.93	(0.86, 1.01)
Age Menarche = Missing	0.85	(0.72, 0.99)	0.89	(0.74, 1.07)	0.99	(0.82, 1.19)	0.94	(0.7, 1.26)
Age FFTP /5	1.03	(0.98, 1.08)	1.02	(0.97, 1.08)	0.99	(0.93, 1.04)	0.94	(0.87, 1.03)
Age FFTP = Missing	7.1	(2.38, 21.12)	0.18	(0.02, 2.04)	15.79	(5.27, 47.26)	7.25	(1.04, 50.64)
Age Menopause	1	(0.99, 1)	0.99	(0.99, 1)	1	(0.99, 1.01)	1	(0.98, 1.01)
Age Menopause = Missing	0.87	(0.72, 1.05)	0.87	(0.7, 1.1)	1.08	(0.88, 1.32)	0.77	(0.5, 1.19)
Alcohol /10	1.05	(1.01, 1.09)	1.06	(1.01, 1.1)	1.01	(0.94, 1.08)	1.02	(0.95, 1.09)
Height /5	1.06	(1.03, 1.09)	1.06	(1.03, 1.1)	1.05	(1.01, 1.08)	1.03	(0.98, 1.09)
Height = Missing	0.73	(0.5, 1.08)	0.69	(0.43, 1.1)	1.41	(0.94, 2.1)	0.61	(0.23, 1.61)
BMI /5	1.18	(1.09, 1.27)	1.15	(1.06, 1.25)	1.02	(0.91, 1.16)	1.1	(0.96, 1.26)
BMI = Missing	0.93	(0.69, 1.25)	1.01	(0.72, 1.41)	0.7	(0.49, 0.99)	0.68	(0.37, 1.28)
Smoke = Never, Former	1	–	1	–	1	–	1	–
Smoke = Current	1.09	(0.99, 1.2)	1.09	(0.98, 1.22)	1.03	(0.9, 1.17)	0.98	(0.83, 1.17)
Smoke = Missing	0.22	(0.08, 0.6)	0.31	(0.08, 1.16)	0.38	(0.14, 1.05)	0	(0, Inf)
PostMeno Never HRT USE	1	–	1	–	1	–	1	–
Premenopause	1.46	(1.26, 1.68)	1.57	(1.34, 1.85)	1.09	(0.92, 1.28)	1.27	(1, 1.63)
PostMeno Former HRT Use	0.94	(0.84, 1.05)	0.95	(0.85, 1.08)	0.9	(0.77, 1.04)	0.9	(0.74, 1.1)
PostMeno Current Non-E HRT Use	1.1	(0.94, 1.3)	1.07	(0.89, 1.3)	0.86	(0.69, 1.06)	0.82	(0.6, 1.13)
PostMeno Current Only E HRT Use	1.27	(1.04, 1.55)	1.28	(1.04, 1.58)	0.79	(0.56, 1.11)	0.73	(0.5, 1.05)
PostMeno Current Any EP HRT Use	1.78	(1.54, 2.06)	1.76	(1.51, 2.06)	1.04	(0.82, 1.32)	0.98	(0.74, 1.29)
Meno/HRT = Missing	0.95	(0.61, 1.46)	0.86	(0.52, 1.43)	0.7	(0.45, 1.1)	1.06	(0.48, 2.33)
(BMI/5)*(PostMeno Never HRT Use)	1	–	1	–	1	–	1	–
(BMI/5)*(Premenopause)	0.72	(0.63, 0.81)	0.73	(0.64, 0.83)	0.87	(0.73, 1.03)	0.83	(0.68, 1)
(BMI/5)*(PostMeno Former HRT Use)	0.97	(0.84, 1.12)	0.97	(0.84, 1.14)	1.06	(0.85, 1.33)	0.88	(0.67, 1.16)
(BMI/5)*(PostMeno Current Non-E HRT Use)	0.75	(0.59, 0.95)	0.77	(0.6, 0.98)	0.69	(0.46, 1.05)	0.64	(0.41, 1)
(BMI/5)*(PostMeno Current Only E HRT Use)	0.71	(0.54, 0.93)	0.73	(0.56, 0.96)	0.83	(0.52, 1.34)	0.67	(0.39, 1.15)
(BMI/5)*(PostMeno Current Any EP HRT Use)	0.85	(0.69, 1.04)	0.88	(0.71, 1.1)	1	(0.73, 1.37)	0.94	(0.63, 1.41)
(BMI/5)*(Meno/HRT = Missing)	1.37	(0.45, 4.22)	1.29	(0.34, 4.92)	0.91	(0.26, 3.25)	0.6	(0.04, 8.98)

"FDA" refers to Full Data Analysis

"CCA" refers to Complete Case Analysis

Figure 4.4: ROC Plot for Models of ER+ and ER- Breast Cancer, Based on Full Data Analysis

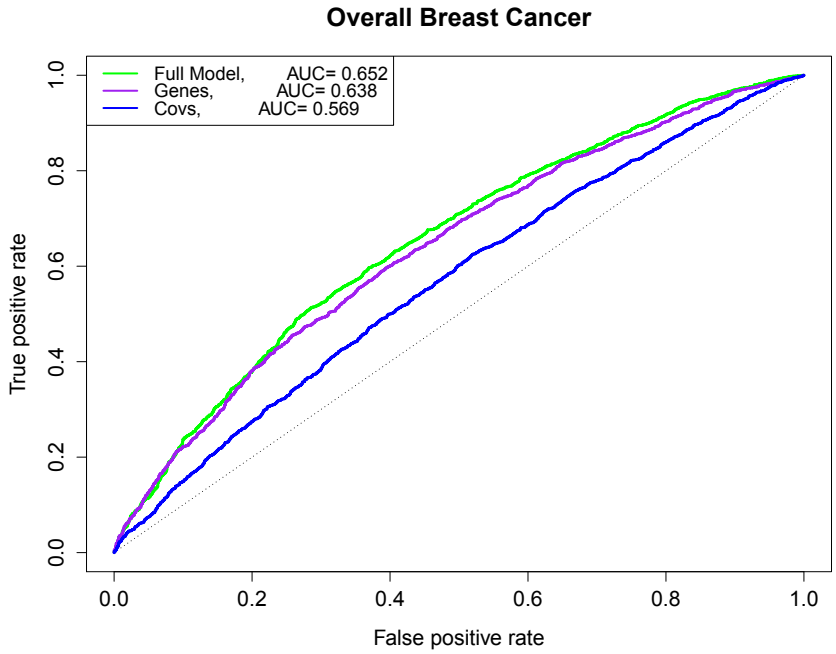


significant, as opposed to only 18 with CCA.

Figure 4.4 presents ROC curves, computed on 2632 controls and 2177 ER+ and 423 ER- cases with complete data, for FDA models with SNPs alone, covariates alone, and the main fully-adjusted model for each subtype. For both subtypes, the SNPs only model provided significantly better discriminatory accuracy than the risk factor only model. The AUC for the ER+ fully-adjusted model was 0.66 while the AUC for the ER- fully-adjusted model was 0.63. Figure 4.5 presents the ROC curves for predicting risk of overall breast cancer based on the sum of the ER+ and ER- risks. The AUC for using subtype-specific risks from the fully-adjusted models to predict overall breast cancer risk based on FDA in the BCAC data is 0.652. This risk discrimination is similar to what we found in the fully-adjusted BPC3 model with 86 SNPs, which had an AUC of 0.654.

Having built the subtype-specific absolute risk models, we projected absolute risk in ages 30-70 for 2,493 controls with no missing data to estimate the distribution of risk in the population for each subtype based on the fully-adjusted model. Figure 4.6 shows the distribution of predicted risks by subtype, along with the proportion of the

Figure 4.5: ROC Plot for Overall Breast Cancer Risk, Defined as the Sum of ER+ and ER- Breast Cancer Risks from Fully-Adjusted Models, Based on Full Data Analysis



population that exceeds certain referent thresholds of risk. Table 4.8 shows the risk stratification provided by the subtype-specific models, giving the average predicted risks within deciles of risk as compared to the average absolute risk in SEER and compared to the average risk for those with and without family history. From an average SEER risk of 7.2% in ages 30-70, the fully-adjusted model for ER+ stratifies risk ranging from average risk of 2.4% to 17.5% in the lowest and highest deciles of risk respectively. The ER- model stratifies risk far less, ranging from 0.8% to 3.7% average risk in the lower and upper deciles of risk, compared to an average SEER risk of 1.8%. Family history alone goes a long way toward that risk stratification, with an average risk of 1.7% for those with no family history and 2.7% for those with family history. The overall risks presented in Table 4.8 are based on the sum of ER+ and ER- risks and result in substantial risk stratification beyond that of family history alone.

Figure 4.6: Distribution of Absolute Risk of Breast Cancer for Ages 30-70 by Estrogen Receptor Status, and the Proportions of the Population with Risk Above Specified Risk Thresholds

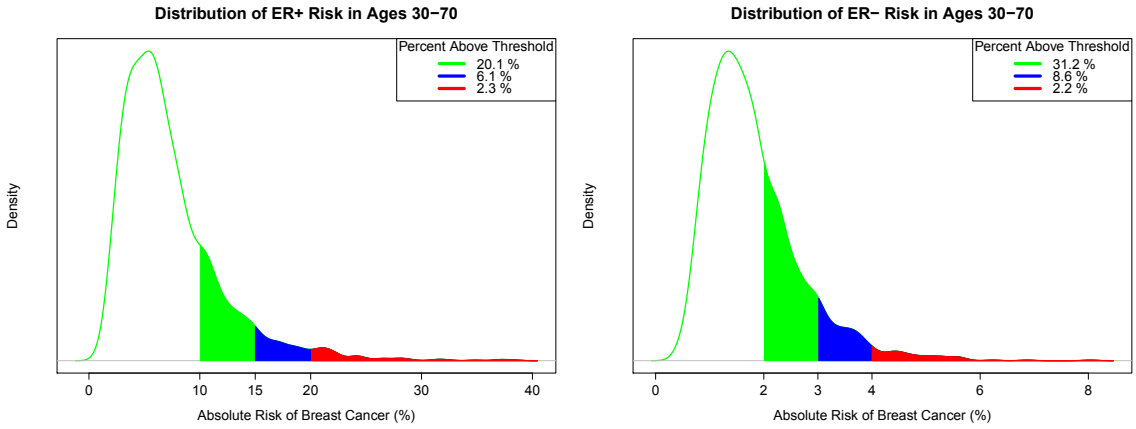


Table 4.8: Average Absolute Risk by Risk Decile for ER Subtypes and Overall Breast Cancer in Ages 30-70

	ER+			ER-			Overall		
SEER	7.2			1.8			8.9		
Fh = No	6.7			1.7			8.4		
Fh = Yes	10.7			2.7			13.4		
Risk Decile	Full Model	Genes	Covs	Full Model	Genes	Covs	Full Model	Genes	Covs
1	2.4	2.9	4.2	0.8	0.9	1.2	3.2	3.9	5.4
2	3.4	3.9	5.1	1.0	1.2	1.4	4.5	5.2	6.6
3	4.2	4.6	5.7	1.2	1.3	1.5	5.5	6.1	7.2
4	5.0	5.3	6.2	1.4	1.5	1.6	6.5	6.8	7.8
5	5.8	6.0	6.7	1.5	1.6	1.7	7.4	7.7	8.5
6	6.6	6.8	7.4	1.7	1.8	1.8	8.4	8.6	9.2
7	7.6	7.6	8.2	1.9	2.0	2.0	9.5	9.6	10.1
8	9.0	8.7	9.1	2.2	2.2	2.1	11.2	10.8	11.1
9	11.2	10.4	10.2	2.6	2.5	2.3	13.7	12.8	12.5
10	17.5	14.4	13.2	3.7	3.2	3.0	20.8	17.4	16.0

Fh = Family History

4.2.4 Conclusions and Future Work

We demonstrated the utility of the methods we developed for building subtype-specific absolute risk models by applying those methods in a real data analysis setting to the BCAC data. We showed how subtype specific absolute risk models can be used to project distributions of risk for each subtype in the population. We found that our absolute risk model for ER+ breast cancer provides substantial risk stratification and thus could be used to target preventative treatments, such as tamoxifen, to women at high risk of ER+ breast cancer. Due to the low rate of ER- breast cancer and a lack of strong known risk factors, our absolute risk model currently produces only modest stratification of risk for ER- breast cancer.

In the future we plan to build off this work and continue to improve the subtype-specific absolute risk models defined by ER status. The absolute risk models we have presented are currently calibrated to the SEER rate for overall breast cancer, with the marginal subtype-specific rates obtained by reweighting according to stratified estimates of the subtype proportions observed in the data. In the future, we plan to instead calibrate directly to the ER stratified rates available in SEER. This requires thinking carefully about how to handle the "ER unknown" SEER rate as improper treatment of this missing ER data can bias results. We plan to use ER+ and ER- rates that correct for the missing ER rate in SEER through a simple published imputation method that accounts for age and calendar year (Anderson et al., 2011).

Additionally, we plan to incorporate survey data, perhaps from the National Health and Nutrition Examination survey (NHANES), in estimation of the attributable risk component of the absolute risk model. Currently, we have estimated the attributable risk for each subtype from cases of that subtype for which there is no missing covariate data, thus basing the estimate on a relatively small amount of data. Even if that criteria were met by large number observations in BCAC, it is still better

to base our estimates of attributable risk on the covariate distribution from NHANES as it is likely to be more representative of the US population than the distribution in the diverse collection of case-control studies in BCAC. Similarly, we also plan to use the covariate distribution from NHANES to specify the set of risk profiles for which to predict risk, which should result in projected risk distributions that are more representative of the US population. Additionally, we will incorporate risks of competing mortality into the model.

We plan to use the BCAC data to build absolute risk models for subtypes that are defined by progesterone receptor (PR) status and human epidermal growth factor receptor 2 (HER2) status as well as ER status. HER2 is an important tumor characteristic for defining breast cancer subtypes as it is used to guide therapy and as a prognostic indicator (Althuis et al., 2004). HER2+ tumors are often associated with more aggressive behavior and worse prognosis than HER2- tumors (Jardines et al., 2005; Burstein, 2005). Treatment by trastuzumab, or Herceptin, is available for HER2+ breast cancers but provides no clinical benefit in HER2-negative breast cancers (Burstein, 2005). Studies of risk factors for breast cancer subtypes defined by HER2 suggest that the risk factors for HER2+ and HER2- tumor types differ as well (Balsari et al., 2003; Yang et al., 2007, 2011b). Given this eterogeneity in risk factors, treatment, and prognosis, an absolute risk model that incorporates HER2 will provide useful information for projecting population risks for these clinically relevant subtypes.

As part of our future work with subtypes defined jointly by ER, PR, and HER2, we plan to develop an absolute risk model for the ER-, PR-, and HER2- subtype, commonly referred to as triple negative breast cancer. Though this subtype is rare, it is clinically relevant due to its biological aggressiveness, lack of treatment options beyond chemotherapy, and poor survival outcomes (Bosch et al., 2010; Cleator et al.,

2007). Women with triple negative breast cancers have worse survival relative to those with other breast cancer subtypes, with a 5-year survival of 77% for the triple negative subtype as compared to 93% for other breast cancers (Bauer et al., 2007). Risk factors for the triple negative subtype have not been well-established, in part because studies may be underpowered to detect associations from small numbers of triple negative cases (Schneider et al., 2008).

The reparametrized multinomial method, which allows cases with missing tumor characteristics or from hospital-based studies to be included in estimation of the hazard ratios, will be vital to building such absolute risk models in the BCAC data as there is a substantial amount of missing data in the PR and HER2 tumor characteristics. The methods we developed for estimating subtype-specific baseline hazard rates will be necessary in this setting, as SEER currently has little data on breast cancer incidence rates stratified by HER2 status. We also plan to validate all absolute risk models that we develop in independent data.

4.3 Supplemental Tables

Supplemental Table 4.1: Percent Completeness of Covariates by Case-Control Status and Study in BPC3

	Overall		CPS2		EPIC		MCCS		MEC	
	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Total	17176	19860	2558	3215	4154	5166	930	765	521	570
Family History	64.3	65.3	97.8	98.7	36.8	29.9	47.6	54.4	89.3	94.4
Menarche	97.8	98.3	98.6	98.9	95.5	96.7	99.8	99.6	99.2	99.8
Parity	97.6	97.7	98.1	98.3	92.6	92.6	100	100	99.0	99.5
Age at First Birth	95.8	91.4	97.7	98.0	95.3	77.4	99.9	100	99.0	98.6
Age at Menopause	87.2	87.6	98.0	98.4	66.2	70.1	86.6	82.7	97.3	98.6
Height	99.7	99.7	99.4	99.4	100	100	100	99.9	99.8	100
Body Mass Index	99.1	98.9	98.9	99.0	100	100	100	99.9	99.8	100
Menopause Status	94.4	94.6	98.5	98.8	82.7	84.3	96.0	100	97.3	98.4
HRT Use: Ever	80.5	80.2	23.9	26.4	75.4	78.5	98.8	99.6	96.7	98.4
HRT Use: Ever E	61.3	60.9	23.2	25.6	44.6	53.8	0	0	95.6	96.5
HRT Use: Ever C	63.6	62.2	23.2	25.6	54.0	59.0	0	0	95.6	96.5
HRT Use: Current	63.0	59.1	0.0	0.0	75.3	68.9	0	0	0	0
Alcohol Use	97.7	97.1	94.5	92.8	100	99.9	100	100	95.8	97.7
Smoking Status	98.9	99.0	98.6	98.9	98.1	98.2	100	100	99.0	99.5
			NHS		PLCO		WHI		WHS	
			Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
Total			1782	3148	790	982	5772	5349	669	665
Family History			99.3	99.6	99.0	99.5	50.5	47.2	96.7	98.2
Menarche			99.2	99.7	100	99.7	97.6	97.8	100	100
Parity			100	100	100	99.8	99.2	99.6	100	100
Age at First Birth			100	100	99.7	99.7	92.1	91.2	99.9	100
Age at Menopause			90.0	91.3	98.9	99.2	94.9	94.7	81.8	77.9
Height			99.9	99.9	99.9	99.7	99.5	99.4	99.4	98.8
Body Mass Index			95.7	96.0	99.6	99.4	99.3	99.2	99.3	98.0
Menopause Status			97.8	97.4	98.9	99.2	100	100	85.2	81.4
HRT Use: Ever			81.5	86.9	99.6	99.2	100	100	95.5	95.8
HRT Use: Ever E			82.1	72.7	0	0	100	100	52.0	45.4
HRT Use: Ever C			82.1	72.7	0	0	100	100	52.0	45.4
HRT Use: Current			63.6	59.6	99.0	97.8	100	100	0	0
Alcohol Use			92.1	93.4	92.7	90.0	99.5	99.6	100	100
Smoking Status			99.9	99.7	100.0	99.9	98.7	99.0	100	99.8

Supplemental Table 4.2: Hazard Ratio Estimates for Fully-Adjusted Model for Overall Breast Cancer with PGRS for Empirical Genotype Data on 24 SNPs in BPC3

	Hazard Ratio	(95% Confidence Interval
PGRS category=1	1	–
PGRS category=2	1.19	(0.76, 1.88)
PGRS category=3	1.33	(0.84, 2.1)
PGRS category=4	1.43	(0.91, 2.25)
PGRS category=5	1.58	(1, 2.48)
PGRS category=6	1.66	(1.06, 2.6)
PGRS category=7	1.78	(1.14, 2.79)
PGRS category=8	2.05	(1.31, 3.21)
PGRS category=9	2.26	(1.45, 3.51)
PGRS decile=10	2.80	(1.82, 4.33)
Family History = No	1	–
Family History = Yes	1.41	(1.02, 1.94)
Menarche category=1	1	–
Menarche category=2	1.07	(0.53, 2.17)
Menarche category=3	0.96	(0.67, 1.38)
Menarche category=4	0.92	(0.65, 1.31)
Menarche category=5	0.90	(0.61, 1.32)
Menarche category=6	0.91	(0.59, 1.41)

Continued on next page...

Supplemental Table 4.2 – continued from previous page

	Hazard Ratios	(95% Confidence Interval)
Menarche category=7	0.86	(0.54, 1.37)
Parity = 0	1	–
Parity = 1	0.84	(0.52, 1.37)
Parity = 2	0.76	(0.49, 1.2)
Parity = 3	0.75	(0.47, 1.18)
Parity = 4	0.72	(0.45, 1.15)
AFFTB category=1	1	–
AFFTB category=2	1.07	(0.69, 1.65)
AFFTB category=3	1.05	(0.65, 1.67)
AFFTB category=4	1.02	(0.68, 1.55)
AFFTB category=5	1.23	(0.78, 1.93)
AFFTB category=6	1.40	(0.87, 2.25)
AFFTB category=7	1.37	(0.85, 2.21)
AFFTB category=8	1.43	(0.76, 2.69)
AFFTB category=9	1.34	(0.58, 3.08)
AgeMeno category=1	1	–
AgeMeno category=2	0.99	(0.64, 1.55)
AgeMeno category=3	0.99	(0.63, 1.56)
AgeMeno category=4	1.06	(0.66, 1.72)

Continued on next page...

Supplemental Table 4.2 – continued from previous page

	Hazard Ratios	(95% Confidence Interval)
AgeMeno category=5	1.15	(0.75, 1.77)
AgeMeno category=6	1.17	(0.72, 1.9)
AgeMeno category=7	1.24	(0.79, 1.95)
AgeMeno category=8	1.30	(0.79, 2.16)
AgeMeno category=9	1.29	(0.8, 2.07)
AgeMeno category=10	1.12	(0.69, 1.84)
Height category=1	1	–
Height category=2	1.12	(0.74, 1.71)
Height category=3	1.14	(0.74, 1.75)
Height category=4	1.17	(0.76, 1.81)
Height category=5	1.10	(0.72, 1.67)
Height category=6	1.23	(0.8, 1.89)
Height category=7	1.20	(0.78, 1.86)
Height category=8	1.31	(0.85, 2.01)
Height category=9	1.22	(0.79, 1.88)
Height category=10	1.31	(0.85, 2.02)
BMI category=1	1	–
BMI category=2	1.06	(0.59, 1.92)
BMI category=3	0.91	(0.49, 1.69)

Continued on next page...

Supplemental Table 4.2 – continued from previous page

	Hazard Ratios	(95% Confidence Interval)
BMI category=4	0.99	(0.53, 1.86)
BMI category=5	0.89	(0.46, 1.71)
BMI category=6	1.09	(0.55, 2.17)
BMI category=7	1.00	(0.49, 2.02)
BMI category=8	0.98	(0.48, 2)
BMI category=9	0.95	(0.46, 1.96)
BMI category=10	1.06	(0.51, 2.21)
PreMeno	1	–
PostMeno HRT=Never	0.59	(0.32, 1.11)
PostMeno Ever HRT	0.68	(0.37, 1.27)
PostMeno Ever E HRT	1.02	(0.7, 1.49)
PostMeno Ever EP HRT	1.28	(0.88, 1.87)
PostMeno HRT=Current	1.24	(0.85, 1.8)
Alcohol category=1	1	–
Alcohol category=2	0.96	(0.65, 1.42)
Alcohol category=3	1.02	(0.68, 1.53)
Alcohol category=4	1.06	(0.72, 1.57)
Alcohol category=5	0.98	(0.66, 1.45)
Alcohol category=6	1.07	(0.72, 1.59)

Continued on next page...

Supplemental Table 4.2 – continued from previous page

	Hazard Ratios	(95% Confidence Interval)
Alcohol category=7	1.18	(0.79, 1.75)
Alcohol category=8	1.25	(0.84, 1.86)
Smoking = Never	1	–
Smoking = Former	1.08	(0.79, 1.47)
Smoking = Current	1.15	(0.8, 1.67)
(BMI category)*(PreMeno)	1	–
(BMI category=2)*(PostMeno HRT=Never)	1.00	(0.5, 2)
(BMI category=3)*(PostMeno HRT=Never)	1.21	(0.6, 2.46)
(BMI category=4)*(PostMeno HRT=Never)	1.20	(0.59, 2.44)
(BMI category=5)*(PostMeno HRT=Never)	1.37	(0.66, 2.83)
(BMI category=6)*(PostMeno HRT=Never)	1.05	(0.49, 2.23)
(BMI category=7)*(PostMeno HRT=Never)	1.33	(0.62, 2.87)
(BMI category=8)*(PostMeno HRT=Never)	1.44	(0.66, 3.12)
(BMI category=9)*(PostMeno HRT=Never)	1.56	(0.71, 3.43)
(BMI category=10)*(PostMeno HRT=Never)	1.46	(0.66, 3.23)
(BMI category=2)*(PostMeno HRT=Ever)	0.96	(0.5, 1.82)
(BMI category=3)*(PostMeno HRT=Ever)	1.11	(0.57, 2.17)
(BMI category=4)*(PostMeno HRT=Ever)	1.12	(0.57, 2.2)
(BMI category=5)*(PostMeno HRT=Ever)	1.18	(0.58, 2.37)

Continued on next page...

Supplemental Table 4.2 – continued from previous page

	Hazard Ratios (95% Confidence Interval)	
(BMI category=6)*(PostMeno HRT=Ever)	1.06	(0.51, 2.19)
(BMI category=7)*(PostMeno HRT=Ever)	1.13	(0.54, 2.4)
(BMI category=8)*(PostMeno HRT=Ever)	1.27	(0.6, 2.71)
(BMI category=9)*(PostMeno HRT=Ever)	1.27	(0.59, 2.74)
(BMI category=10)*(PostMeno HRT=Ever)	1.25	(0.57, 2.74)

Supplemental Table 4.3: Hazard Ratios (HR) and 95% Confidence Intervals (CI) for SNPs in Fully-Adjusted Models of ER+ and ER- Breast Cancer, with Results from Full Data Analysis (FDA) and Complete Case Analysis (CCA)

	ER+				ER-			
	FDA		CCA		FDA		CCA	
SNP	HR	(CI)	HR	(CI)	HR	(CI)	HR	(CI)
rs78540526	1.38	(1.04, 1.83)	1.38	(1.02, 1.86)	0.79	(0.55, 1.15)	0.74	(0.39, 1.42)
rs75915166	1.26	(1.05, 1.5)	1.27	(1.04, 1.56)	1.31	(1.08, 1.59)	0.9	(0.64, 1.26)
rs554219	1.23	(1.15, 1.32)	1.24	(1.15, 1.34)	1.19	(1.1, 1.29)	1.23	(1.08, 1.39)
rs7726159	1.2	(1, 1.45)	1.23	(1.01, 1.5)	1.09	(0.89, 1.35)	1.27	(0.93, 1.73)
rs10069690	1.18	(1.06, 1.33)	1.2	(1.05, 1.36)	1.07	(0.94, 1.22)	0.97	(0.79, 1.19)
rs2736108	1.15	(1.09, 1.21)	1.16	(1.1, 1.23)	1.06	(1, 1.13)	1.03	(0.94, 1.12)
rs2588809	1.14	(1.06, 1.23)	1.16	(1.04, 1.29)	1.02	(0.94, 1.11)	1.15	(0.98, 1.36)

Continued on next page...

Supplemental Table 4.3 – continued from previous page

	ER+				ER-			
	FDA		CCA		FDA		CCA	
SNP	HR	(CI)	HR	(CI)	HR	(CI)	HR	(CI)
rs999737	1.13	(1.03, 1.24)	1.15	(1.08, 1.23)	1.13	(1.02, 1.25)	1.05	(0.95, 1.16)
rs10759243	1.13	(1.07, 1.19)	1.15	(1.04, 1.27)	1.09	(1.03, 1.16)	1.1	(0.95, 1.27)
rs865686	1.13	(1.03, 1.22)	1.11	(1.05, 1.18)	1.13	(1.03, 1.24)	0.95	(0.87, 1.04)
rs2981579	1.11	(1.05, 1.18)	1.11	(1.04, 1.18)	1.05	(0.98, 1.13)	1.03	(0.93, 1.13)
rs11199914	1.11	(1.05, 1.17)	1.1	(1.01, 1.2)	1.02	(0.95, 1.08)	1.04	(0.91, 1.2)
rs7072776	1.11	(0.97, 1.26)	1.1	(0.98, 1.23)	1.05	(0.9, 1.22)	1.2	(1.02, 1.42)
rs11814448	1.1	(1.04, 1.16)	1.09	(1, 1.2)	1.03	(0.97, 1.09)	1.1	(0.95, 1.26)
rs13387042	1.1	(1.01, 1.19)	1.09	(1.02, 1.17)	1.1	(1.01, 1.2)	1	(0.9, 1.11)
rs16857609	1.09	(1.04, 1.15)	1.09	(1.02, 1.16)	1.02	(0.96, 1.08)	1	(0.91, 1.1)
rs11552449	1.09	(1.03, 1.16)	1.09	(1.02, 1.16)	1.14	(1.07, 1.22)	1.09	(0.99, 1.19)
rs11249433	1.09	(1.04, 1.14)	1.09	(1.02, 1.16)	1.06	(1.01, 1.12)	0.99	(0.9, 1.09)
rs1045485	1.09	(0.98, 1.2)	1.08	(1.01, 1.16)	1.12	(1.01, 1.25)	0.96	(0.87, 1.07)
rs4973768	1.08	(1.02, 1.14)	1.08	(1.02, 1.14)	1.05	(0.98, 1.11)	1.05	(0.96, 1.14)
rs10941679	1.08	(1.02, 1.14)	1.08	(1.01, 1.16)	1.02	(0.96, 1.09)	1.14	(1.02, 1.27)
rs889312	1.07	(1.02, 1.13)	1.08	(1.01, 1.15)	1.05	(0.99, 1.11)	0.96	(0.87, 1.07)
rs12662670	1.07	(1.01, 1.14)	1.08	(1.01, 1.15)	1.11	(1.04, 1.19)	1.21	(1.09, 1.34)
rs2046210	1.07	(1.01, 1.13)	1.08	(1.02, 1.14)	1.03	(0.97, 1.09)	1.02	(0.94, 1.11)

Continued on next page...

Supplemental Table 4.3 – continued from previous page

	ER+				ER-			
	FDA		CCA		FDA		CCA	
SNP	HR	(CI)	HR	(CI)	HR	(CI)	HR	(CI)
rs13281615	1.06	(1.01, 1.12)	1.08	(1.01, 1.14)	1.03	(0.98, 1.09)	0.95	(0.87, 1.05)
rs1011970	1.06	(1, 1.13)	1.07	(1.01, 1.13)	1.08	(1.01, 1.16)	1.12	(1.03, 1.23)
rs2380205	1.06	(1.01, 1.12)	1.07	(1.01, 1.14)	1.03	(0.97, 1.1)	0.96	(0.87, 1.05)
rs10995190	1.06	(1, 1.13)	1.07	(1.01, 1.13)	1.02	(0.95, 1.09)	1.02	(0.94, 1.12)
rs704010	1.06	(1.01, 1.12)	1.07	(1, 1.14)	1.01	(0.96, 1.07)	1.03	(0.94, 1.14)
rs3817198	1.06	(1.01, 1.11)	1.07	(1.01, 1.13)	1.1	(1.04, 1.16)	0.97	(0.89, 1.06)
rs10771399	1.06	(1.01, 1.11)	1.06	(1, 1.12)	1.06	(1, 1.12)	1.07	(0.98, 1.16)
rs1292011	1.06	(1.01, 1.11)	1.06	(1, 1.12)	1	(0.95, 1.06)	1.05	(0.96, 1.14)
rs3803662	1.05	(1, 1.11)	1.06	(1, 1.12)	1.02	(0.96, 1.08)	1.04	(0.95, 1.14)
rs6504950	1.05	(1, 1.1)	1.06	(0.91, 1.23)	1.02	(0.97, 1.08)	1.12	(0.89, 1.41)
rs8170	1.05	(0.99, 1.11)	1.05	(0.99, 1.13)	1.03	(0.96, 1.1)	1.13	(1.02, 1.25)
rs2363956	1.03	(0.98, 1.09)	1.05	(0.96, 1.15)	1.03	(0.98, 1.09)	0.84	(0.74, 0.97)
rs2823093	1.03	(0.98, 1.08)	1.04	(0.98, 1.1)	1.05	(0.99, 1.11)	1.06	(0.97, 1.16)
rs17879961	1.03	(0.98, 1.08)	1.03	(0.95, 1.12)	0.95	(0.9, 1.01)	0.96	(0.84, 1.08)
rs616488	1.01	(0.94, 1.09)	1.02	(0.96, 1.08)	0.97	(0.89, 1.05)	1.04	(0.95, 1.14)
rs4849887	0.99	(0.94, 1.04)	1.02	(0.95, 1.08)	1.05	(0.99, 1.11)	0.88	(0.79, 0.97)
rs2016394	0.99	(0.93, 1.05)	1.01	(0.96, 1.07)	0.93	(0.87, 0.99)	1	(0.92, 1.1)

Continued on next page...

Supplemental Table 4.3 – continued from previous page

	ER+				ER-			
	FDA		CCA		FDA		CCA	
SNP	HR	(CI)	HR	(CI)	HR	(CI)	HR	(CI)
rs1550623	0.98	(0.91, 1.06)	1	(0.74, 1.37)	0.98	(0.9, 1.06)	0.63	(0.42, 0.94)
rs6762644	0.98	(0.91, 1.06)	0.99	(0.82, 1.19)	0.94	(0.86, 1.03)	0.79	(0.58, 1.07)
rs12493607	0.97	(0.92, 1.02)	0.99	(0.93, 1.05)	0.97	(0.92, 1.03)	1.07	(0.98, 1.17)
rs9790517	0.97	(0.92, 1.02)	0.99	(0.93, 1.04)	0.95	(0.9, 1.01)	0.95	(0.87, 1.05)
rs6828523	0.97	(0.92, 1.02)	0.97	(0.91, 1.04)	0.98	(0.92, 1.04)	0.94	(0.85, 1.04)
rs10472076	0.97	(0.91, 1.02)	0.97	(0.91, 1.03)	1.06	(1, 1.13)	0.98	(0.89, 1.08)
rs1353747	0.96	(0.92, 1.01)	0.97	(0.9, 1.04)	0.95	(0.9, 1.01)	0.85	(0.76, 0.95)
rs1432679	0.96	(0.91, 1.02)	0.97	(0.91, 1.03)	1.01	(0.95, 1.08)	0.96	(0.88, 1.05)
rs11242675	0.96	(0.88, 1.05)	0.97	(0.91, 1.03)	0.9	(0.82, 1)	1.03	(0.94, 1.13)
rs204247	0.96	(0.9, 1.02)	0.96	(0.9, 1.03)	0.97	(0.91, 1.04)	1.07	(0.97, 1.18)
rs17529111	0.96	(0.89, 1.03)	0.96	(0.9, 1.02)	0.94	(0.87, 1.02)	0.96	(0.87, 1.06)
rs720475	0.95	(0.89, 1.02)	0.96	(0.9, 1.01)	0.99	(0.92, 1.07)	0.87	(0.8, 0.96)
rs9693444	0.95	(0.73, 1.24)	0.95	(0.9, 1.01)	0.91	(0.68, 1.21)	0.94	(0.86, 1.03)
rs6472903	0.95	(0.9, 1.01)	0.95	(0.88, 1.04)	0.95	(0.89, 1.01)	0.93	(0.81, 1.06)
rs2943559	0.95	(0.8, 1.12)	0.95	(0.9, 1.01)	0.93	(0.77, 1.12)	0.86	(0.79, 0.94)
rs11780156	0.94	(0.88, 1.01)	0.95	(0.89, 1.01)	0.92	(0.85, 0.99)	0.9	(0.82, 0.99)
rs7904519	0.94	(0.89, 0.99)	0.94	(0.87, 1.03)	0.9	(0.85, 0.96)	1.05	(0.92, 1.19)

Continued on next page...

Supplemental Table 4.3 – continued from previous page

	ER+				ER-			
	FDA		CCA		FDA		CCA	
SNP	HR	(CI)	HR	(CI)	HR	(CI)	HR	(CI)
rs3903072	0.94	(0.89, 0.99)	0.94	(0.88, 1.02)	0.93	(0.88, 0.99)	1.04	(0.93, 1.18)
rs11820646	0.94	(0.89, 0.99)	0.92	(0.87, 0.98)	0.86	(0.81, 0.92)	0.89	(0.81, 0.98)
rs12422552	0.93	(0.89, 0.98)	0.92	(0.87, 0.97)	0.96	(0.91, 1.02)	0.91	(0.83, 0.99)
rs17356907	0.93	(0.88, 0.98)	0.92	(0.86, 0.98)	0.87	(0.82, 0.93)	0.93	(0.84, 1.02)
rs11571833	0.92	(0.88, 0.97)	0.92	(0.86, 0.97)	0.97	(0.91, 1.02)	0.88	(0.8, 0.96)
rs2236007	0.92	(0.87, 0.97)	0.91	(0.86, 0.97)	0.95	(0.89, 1.01)	0.98	(0.9, 1.07)
rs941764	0.92	(0.87, 0.97)	0.91	(0.86, 0.96)	0.94	(0.88, 1)	1.04	(0.95, 1.14)
rs17817449	0.92	(0.85, 0.98)	0.91	(0.82, 1)	0.93	(0.86, 1)	1.02	(0.87, 1.19)
rs13329835	0.91	(0.87, 0.97)	0.91	(0.84, 0.98)	0.99	(0.93, 1.06)	1	(0.89, 1.12)
rs527616	0.91	(0.86, 0.98)	0.9	(0.83, 0.98)	0.94	(0.87, 1.01)	0.99	(0.88, 1.12)
rs1436904	0.91	(0.87, 0.96)	0.9	(0.84, 0.96)	0.94	(0.88, 0.99)	1.1	(1, 1.22)
rs4808801	0.91	(0.85, 0.97)	0.89	(0.85, 0.95)	0.96	(0.89, 1.03)	0.97	(0.89, 1.06)
rs3760982	0.9	(0.86, 0.95)	0.88	(0.82, 0.95)	0.96	(0.9, 1.01)	1	(0.89, 1.13)
rs132390	0.9	(0.84, 0.96)	0.88	(0.83, 0.93)	0.94	(0.87, 1.01)	0.99	(0.91, 1.08)
rs6001930	0.89	(0.85, 0.94)	0.87	(0.81, 0.94)	0.95	(0.9, 1.01)	0.97	(0.86, 1.08)
rs4245739	0.87	(0.83, 0.92)	0.86	(0.79, 0.94)	0.94	(0.89, 0.99)	0.77	(0.67, 0.89)
rs6678914	0.82	(0.76, 0.89)	0.85	(0.81, 0.91)	0.79	(0.72, 0.86)	0.92	(0.84, 1)

Continued on next page...

Supplemental Table 4.3 – continued from previous page

	ER+				ER-			
	FDA		CCA		FDA		CCA	
SNP	HR	(CI)	HR	(CI)	HR	(CI)	HR	(CI)
rs12710696	0.8	(0.76, 0.85)	0.8	(0.75, 0.85)	0.84	(0.79, 0.89)	0.83	(0.75, 0.91)
rs11075995	0.76	(0.73, 0.8)	0.74	(0.7, 0.78)	0.91	(0.86, 0.96)	0.97	(0.89, 1.06)

Bibliography

MB Adamo, CH Johnson, JL Ruhl, and LA Dickie. Seer program coding and staging manual 2011, appendix c. *National Cancer Institute*, 2011. URL `http://seer.cancer.gov/tools/codingmanuals/index.html`.

Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, 2002.

Michelle D Althuis, Jennifer H Fergenbaum, Montserrat Garcia-Closas, Louise A Brinton, M Patricia Madigan, and Mark E Sherman. Etiology of hormone receptor-defined breast cancer: A systematic review of the literature. *Cancer epidemiology biomarkers prevention a publication of the American Association for Cancer Research cosponsored by the American Society of Preventive Oncology*, 13(10):1558–1568, 2004. URL `http://www.ncbi.nlm.nih.gov/pubmed/15466970`.

W. F. Anderson, H. A. Katki, and P. S. Rosenberg. Incidence of breast cancer in the United States: current and future trends. *J. Natl. Cancer Inst.*, 103(18):1397–1402, Sep 2011.

William F Anderson and Rayna Matsuno. Breast cancer heterogeneity: A mixture of at least two main types? *Journal of the National Cancer Institute*, 98(14):948–951, 2006. URL `http://www.ncbi.nlm.nih.gov/pubmed/16849671`.

O. G. Bahcall. iCOGS collection provides a collaborative model. Foreword. *Nat. Genet.*, 45(4):343, Apr 2013.

A Balsari, P Casalini, R Bufalino, F Berrino, and S Mnard. Role of hormonal risk factors in her2-positive breast carcinomas. *British Journal of Cancer*, 88(7):1032–1034, 2003. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2376362&tool=pmcentrez&rendertype=abstract`.

Katrina R. Bauer, Monica Brown, Rosemary D. Cress, Carol A. Parise, and Vincent Caggiano. Descriptive analysis of estrogen receptor (er)-negative, progesterone receptor (pr)-negative, and her2-negative invasive breast cancer, the so-called triple-negative phenotype. *Cancer*, 109(9):1721–1728, 2007. ISSN 1097-0142. doi: 10.1002/cncr.22618. URL `http://dx.doi.org/10.1002/cncr.22618`.

A. Bosch, P. Eroles, R. Zaragoza, J. R. Vina, and A. Lluch. Triple-negative breast cancer: molecular features, pathogenesis, treatment and current lines of research. *Cancer Treat. Rev.*, 36(3):206–215, May 2010.

Breast Cancer Association Consortium. Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. *J. Natl. Cancer Inst.*, 98(19):1382–1396, Oct 2006.

NE. Breslow. Contribution to the discussion of paper by D.R. Cox. *J R Stat Soc Ser B*, 34:216–217, 1972.

P Bruzzi, S B Green, D P Byar, L A Brinton, and C Schairer. Estimating the population attributable risk for multiple risk factors using case-control data. *American Journal of Epidemiology*, 122(5):904–914, 1985. URL `http://www.ncbi.nlm.nih.gov/pubmed/4050778`.

Harold J Burstein. The distinctive nature of her2-positive breast cancers. *The New England Journal of Medicine*, 353(16):1652–1654, 2005. URL `http://www.ncbi.nlm.nih.gov/pubmed/16236735`.

C. D. Bushnell and L. B. Goldstein. Risk of ischemic stroke with tamoxifen treatment for breast cancer: a meta-analysis. *Neurology*, 63(7):1230–1233, Oct 2004.

N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J. H. Park. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.*, 45(4):400–405, Apr 2013.

Yi-Hau Chen and Hung Chen. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(3):pp. 449–460, 2000. ISSN 13697412. URL `http://www.jstor.org/stable/2680690`.

R. T. Chlebowski, N. Col, E. P. Winer, D. E. Collyar, S. R. Cummings, V. G. Vogel, H. J. Burstein, A. Eisen, I. Lipkus, and D. G. Pfister. American Society of Clinical Oncology technology assessment of pharmacologic interventions for breast cancer risk reduction including tamoxifen, raloxifene, and aromatase inhibition. *J. Clin. Oncol.*, 20(15):3328–3343, Aug 2002.

S. Cleator, W. Heller, and R. C. Coombes. Triple-negative breast cancer: therapeutic options. *Lancet Oncol.*, 8(3):235–244, Mar 2007.

G. A. Colditz, J. E. Manson, and S. E. Hankinson. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J Womens Health*, 6 (1):49–62, Feb 1997.

Amy Berrington de Gonzalez, A Iulian Apostoaei, Lene H S Veiga, Preetha Rajaraman, Brian A Thomas, F Owen Hoffman, Ethel Gilbert, and Charles Land. Radrat:

a radiation risk assessment tool for lifetime cancer risk projection. *Journal of Radiological Protection*, 32(3):205, 2012. URL `http://stacks.iop.org/0952-4746/32/i=3/a=205`.

R. M. Elledge, G. M. Clark, G. C. Chamness, and C. K. Osborne. Tumor biologic factors and breast cancer prognosis among white, Hispanic, and black women in the United States. *J. Natl. Cancer Inst.*, 86(9):705–712, May 1994.

Peter A. Fasching, Paul D.P. Pharoah, Angela Cox, Heli Nevanlinna, Stig E. Bojesen, Thomas Karn, Annegien Broeks, Flora E. van Leeuwen, Laura J. van 't Veer, Renate Udo, Alison M. Dunning, Dario Greco, Kristiina Aittomki, Carl Blomqvist, Mitul Shah, Brge G. Nordestgaard, Henrik Flyger, John L. Hopper, Melissa C. Southey, Carmel Apicella, Montserrat Garcia-Closas, Mark Sherman, Jolanta Lissowska, Caroline Seynaeve, Petra E.A. Huijts, Rob A.E.M. Tollenaar, Argyrios Ziogas, Arif B. Ekici, Claudia Rauh, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M. Hartikainen, Irene L. Andrulis, Hilmi Ozcelik, Anna-Marie Mulligan, Gord Glendon, Per Hall, Kamila Czene, Jianjun Liu, Jenny Chang-Claude, Shan Wang-Gohrke, Ursula Eilber, Stefan Nickels, Thilo Drk, Maria Schiekel, Michael Bremer, Tjoung-Won Park-Simon, Graham G. Giles, Gianluca Severi, Laura Baglietto, Maartje J. Hooning, John W.M. Martens, Agnes Jager, Mieke Kriege, Annika Lindblom, Sara Margolin, Fergus J. Couch, Kristen N. Stevens, Janet E. Olson, Matthew Kosel, Simon S. Cross, Sabapathy P. Balasubramanian, Malcolm W.R. Reed, Alexander Miron, Esther M. John, Robert Winqvist, Katri Pylks, Arja Jukkola-Vuorinen, Saila Kauppila, Barbara Burwinkel, Frederik Marme, Andreas Schneeweiss, Christof Sohn, Georgia Chenevix-Trench, kConFab Investigators, Diether Lambrechts, Anne-Sophie Dieudonne, Sigrid Hatse,

Erik van Limbergen, Javier Benitez, Roger L. Milne, M. Pilar Zamora, Jos Ignacio Arias Prez, Bernardo Bonanni, Bernard Peissel, Bernard Loris, Paolo Peterlongo, Preetha Rajaraman, Sara J. Schonfeld, Hoda Anton-Culver, Peter Devilee, Matthias W. Beckmann, Dennis J. Slamon, Kelly-Anne Phillips, Jonine D. Figueroa, Manjeet K. Humphreys, Douglas F. Easton, and Marjanka K. Schmidt. The role of genetic breast cancer susceptibility variants as prognostic factors. *Human Molecular Genetics*, 21(17):3926–3939, 2012. doi: 10.1093/hmg/dds159. URL `http://hmg.oxfordjournals.org/content/21/17/3926.abstract`.

Bernard Fisher, Joseph P. Costantino, D. Lawrence Wickerham, Carol K. Redmond, Maureen Kavanah, Walter M. Cronin, Victor Vogel, Andr Robidoux, Nikolay Dimitrov, James Atkins, Mary Daly, Samuel Wieand, Elizabeth Tan-Chiu, Leslie Ford, Norman Wolmark, other National Surgical Adjuvant Breast, and Bowel Project Investigators. Tamoxifen for prevention of breast cancer: Report of the national surgical adjuvant breast and bowel project p-1 study. *Journal of the National Cancer Institute*, 90(18):1371–1388, 1998. doi: 10.1093/jnci/90.18.1371. URL `http://jnci.oxfordjournals.org/content/90/18/1371.abstract`.

A. N. Freedman, M. L. Slattery, R. Ballard-Barbash, G. Willis, B. J. Cann, D. Pee, M. H. Gail, and R. M. Pfeiffer. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *J. Clin. Oncol.*, 27(5):686–693, Feb 2009a.

A. N. Freedman, M. L. Slattery, R. Ballard-Barbash, G. Willis, B. J. Cann, D. Pee, M. H. Gail, and R. M. Pfeiffer. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *J. Clin. Oncol.*, 27(5):686–693, Feb 2009b.

A Fritz and L Ries. Seer program code manual, 3rd edition. *National Cancer Institute*, 1998. URL http://seer.cancer.gov/manuals/historic/codeman.pdf.

MITCHELL H. Gail. The estimation and use of absolute risk for weighing the risks and benefits of selective estrogen receptor modulators for preventing breast cancer. *Annals of the New York Academy of Sciences*, 949(1):286–291, 2001. ISSN 1749-6632. doi: 10.1111/j.1749-6632.2001.tb04034.x. URL http://dx.doi.org/10.1111/j.1749-6632.2001.tb04034.x.

Mitchell H Gail. Personalized Estimates of Breast Cancer Risk in Clinical Practice and Public Health. *Statistics in Medicine*, 30(10):1090–1104, 2011. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3079423\&tool=pmcentrez\&rendertype=abstract.

Mitchell H Gail, Louise A Brinton, David P Byar, K Donald, Sylvan B Green, Catherine Schairer, and John J Mutvihill. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *Journal Of The National Cancer Institute*, pages 1879–1886, 1989.

B. I. Graubard and T. R. Fears. Standard errors for attributable risk for simple and complex sample designs. *Biometrics*, 61(3):847–855, Sep 2005.

S. M. Grundy. Primary prevention of coronary heart disease: selection of patients for aggressive cholesterol management. *Am. J. Med.*, 107(2A):2S–6S, Aug 1999.

Michael T. Halpern, Brenda W. Gillespie, and Kenneth E. Warner. Patterns of absolute risk of lung cancer mortality in former smokers. *Journal of the National Cancer Institute*, 85(6):457–464, 1993. doi: 10.1093/jnci/85.6.457. URL http://jnci.oxfordjournals.org/content/85/6/457.abstract.

J. A. Hanley. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging*, 29(3):307–335, 1989.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

J D Holt. Competing risk analyses with special reference to matched pair experiments. *Biometrika*, 65(1):159–165, 1978. URL `http://biomet.oxfordjournals.org/cgi/content/abstract/65/1/159`.

N Howlader, AM Noone, M Krapcho, N Neyman, R Aminou, W Waldron, CL Altekruse, SF amd Kosary, J Ruhl, Z Tatalovich, H Cho, and et al. Seer cancer statistics review, 1975-2008. *National Cancer Institute*, 2011. URL `http://seer.cancer.gov/csr/1975_2008/`.

D. J. Hunter, E. Riboli, C. A. Haiman, D. Albanes, D. Altshuler, S. J. Chanock, R. B. Haynes, B. E. Henderson, R. Kaaks, D. O. Stram, G. Thomas, M. J. Thun, H. Blanche, J. E. Buring, N. P. Burtt, E. E. Calle, H. Cann, F. Canzian, Y. C. Chen, G. A. Colditz, D. G. Cox, A. M. Dunning, H. S. Feigelson, M. L. Freedman, J. M. Gaziano, E. Giovannucci, S. E. Hankinson, J. N. Hirschhorn, R. N. Hoover, T. Key, L. N. Kolonel, P. Kraft, L. Le Marchand, S. Liu, J. Ma, S. Melnick, P. Pharaoh, M. C. Pike, C. Rodriguez, V. W. Setiawan, M. J. Stampfer, E. Trapido, R. Travis, J. Virtamo, S. Wacholder, and W. C. Willett. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nat. Rev. Cancer*, 5(12):977–985, Dec 2005.

A. Husing, F. Canzian, L. Beckmann, M. Garcia-Closas, W. R. Diver, M. J. Thun, C. D. Berg, R. N. Hoover, R. G. Ziegler, J. D. Figueroa, C. Isaacs, A. Olsen,

V. Viallon, H. Boeing, G. Masala, D. Trichopoulos, P. H. Peeters, E. Lund, E. Ardanaz, K. T. Khaw, P. Lenner, L. N. Kolonel, D. O. Stram, L. Le Marchand, C. A. McCarty, J. E. Buring, I. M. Lee, S. Zhang, S. Lindstrom, S. E. Hankinson, E. Riboli, D. J. Hunter, B. E. Henderson, S. J. Chanock, C. A. Haiman, P. Kraft, and R. Kaaks. Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. *J. Med. Genet.*, 49(9):601–608, Sep 2012.

National Cancer Institute. Current bpc3 research plan, May 2014. URL `http://epi.grants.cancer.gov/BPC3/abstract2.html`.

R. Jackson. Guidelines on preventing cardiovascular disease in clinical practice. *BMJ*, 320(7236):659–661, Mar 2000.

R. Jackson, C. M. Lawes, D. A. Bennett, R. J. Milne, and A. Rodgers. Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. *Lancet*, 365(9457):434–441, 2005.

Lori Jardines, Bruce G Haffty, Paul Fisher, Jeffrey Weitzel, Melanie Royce, and P D. Breast cancer overview: Risk factors, screening, genetic testing, and prevention. *Oncology*, 2005.

Ismail Jatoi, Bingshu E. Chen, William F. Anderson, and Philip S. Rosenberg. Breast cancer mortality trends in the united states according to estrogen receptor status and age at diagnosis. *Journal of Clinical Oncology*, 25(13):1683–1690, 2007. doi: 10.1200/JCO.2006.09.2106. URL `http://jco.ascopubs.org/content/25/13/1683.abstract`.

J D Kalbfleisch and R L Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, 1980. URL `http://proquest.umi.com/pqdweb?did=745641091&Fmt=7&clientId=3748&RQT=309&VName=PQD`.

P. H. Lahmann, K. Hoffmann, N. Allen, C. H. van Gils, K. T. Khaw, B. Tehard, F. Berrino, A. Tj?nneland, J. Bigaard, A. Olsen, K. Overvad, F. Clavel-Chapelon, G. Nagel, H. Boeing, D. Trichopoulos, G. Economou, G. Bellos, D. Palli, R. Tumino, S. Panico, C. Sacerdote, V. Krogh, P. H. Peeters, H. B. Bueno-de Mesquita, E. Lund, E. Ardanaz, P. Amiano, G. Pera, J. R. Quiros, C. Martinez, M. J. Tormo, E. Wirfalt, G. Berglund, G. Hallmans, T. J. Key, G. Reeves, S. Bingham, T. Norat, C. Biessy, R. Kaaks, and E. Riboli. Body size and breast cancer risk: findings from the European Prospective Investigation into Cancer And Nutrition (EPIC). *Int. J. Cancer*, 111(5):762–771, Sep 2004.

Petra H. Lahmann, Lauren Lissner, Bo Gullberg, Hkan Olsson, and Gran Berglund. A prospective study of adiposity and postmenopausal breast cancer risk: The malm diet and cancer study. *International Journal of Cancer*, 103(2):246–252, 2003. ISSN 1097-0215. doi: 10.1002/ijc.10799. URL `http://dx.doi.org/10.1002/ijc.10799`.

H. Ma, L. Bernstein, M. C. Pike, and G. Ursin. Reproductive factors and breast cancer risk according to joint estrogen and progesterone receptor status: a meta-analysis of epidemiological studies. *Breast Cancer Res.*, 8(4):R43, 2006.

R. L. Milne, M. M. Gaudet, A. B. Spurdle, P. A. Fasching, F. J. Couch, J. Benitez, J. I. Arias Perez, M. P. Zamora, N. Malats, I. Dos Santos Silva, L. J. Gibson, O. Fletcher, N. Johnson, H. Anton-Culver, A. Ziogas, J. Figueroa, L. Brinton, M. E. Sherman, J. Lissowska, J. L. Hopper, G. S. Dite, C. Apicella, M. C. Southey, A. J. Sigurdson, M. S. Linet, S. J. Schonfeld, D. M. Freedman, A. Mannermaa, V. M. Kosma, V. Kataja, P. Auvinen, I. L. Andrulis, G. Glendon, J. A. Knight, N. Weerasooriya, A. Cox, M. W. Reed, S. S. Cross, A. M. Dunning, S. Ahmed, M. Shah, H. Brauch, Y. D. Ko, T. Bruning, D. Lambrechts, J. Reumers, A. Smeets,

S. Wang-Gohrke, P. Hall, K. Czene, J. Liu, A. K. Irwanto, G. Chenevix-Trench, H. Holland, G. G. Giles, L. Baglietto, G. Severi, S. E. Bojensen, B. G. Nordestgaard, H. Flyger, E. M. John, D. W. West, A. S. Whittemore, C. Vachon, J. E. Olson, Z. Fredericksen, M. Kosel, R. Hein, A. Vrieling, D. Flesch-Janys, J. Heinz, M. W. Beckmann, K. Heusinger, A. B. Ekici, L. Haeberle, M. K. Humphreys, J. Morrison, D. F. Easton, P. D. Pharoah, M. Garcia-Closas, E. L. Goode, and J. Chang-Claude. Assessing interactions between the associations of common genetic susceptibility variants, reproductive history and body mass index with breast cancer risk in the breast cancer association consortium: a combined case-control study. *Breast Cancer Res.*, 12(6):R110, 2010.

L. M. Morimoto, E. White, Z. Chen, R. T. Chlebowski, J. Hays, L. Kuller, A. M. Lopez, J. Manson, K. L. Margolis, P. C. Muti, M. L. Stefanick, and A. McTiernan. Obesity, body size, and risk of postmenopausal breast cancer: the Women's Health Initiative (United States). *Cancer Causes Control*, 13(8):741–751, Oct 2002.

V. A. Moyer, V. A. Moyer, M. L. LeFevre, A. L. Siu, J. J. Peters, J. J. Baumann, K. Bibbins-Domingo, S. J. Curry, M. Ebell, G. Flores, F. A. Garc?a, A. G. Cantu, D. C. Grossman, J. Herzstein, W. K. Nicholson, D. K. Owens, W. R. Phillips, and M. P. Pignone. Medications to decrease the risk for breast cancer in women: recommendations from the U.S. Preventive Services Task Force recommendation statement. *Ann. Intern. Med.*, 159(10):698–708, Nov 2013.

C. J. Murray, J. A. Lauer, R. C. Hutubessy, L. Niessen, N. Tomijima, A. Rodgers, C. M. Lawes, and D. B. Evans. Effectiveness and costs of interventions to lower systolic blood pressure and cholesterol: a global and regional analysis on reduction of cardiovascular-disease risk. *Lancet*, 361(9359):717–725, Mar 2003.

Stefan Nickels, Thrse Truong, Rebecca Hein, Kristen Stevens, Katharina Buck, Sabine

Behrens, Ursula Eilber, Martina Schmidt, Lothar Hberle, Alina Vrieling, Mia Gaudet, Jonine Figueroa, Nils Schoof, Amanda B. Spurdle, Anja Rudolph, Peter A. Fasching, John L. Hopper, Enes Makalic, Daniel F. Schmidt, Melissa C. Southey, Matthias W. Beckmann, Arif B. Ekici, Olivia Fletcher, Lorna Gibson, Isabel dos Santos Silva, Julian Peto, Manjeet K. Humphreys, Jean Wang, Emilie Cordina-Duverger, Florence Menegaux, Brge G. Nordestgaard, Stig E. Bojesen, Charlotte Lanng, Hoda Anton-Culver, Argyrios Ziogas, Leslie Bernstein, Christina A. Clarke, Hermann Brenner, Heiko Mller, Volker Arndt, Christa Stegmaier, Hiltrud Brauch, Thomas Brning, Volker Harth, The GENICA Network, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M. Hartikainen, kConFab, AOCS Management Group, Diether Lambrechts, Dominiek Smeets, Patrick Neven, Robert Paridaens, Dieter Flesch-Janys, Nadia Obi, Shan Wang-Gohrke, Fergus J. Couch, Janet E. Olson, Celine M. Vachon, Graham G. Giles, Gianluca Severi, Laura Baglietto, Kenneth Offit, Esther M. John, Alexander Miron, Irene L. Andrulis, Julia A. Knight, Gord Glendon, Anna Marie Mulligan, Stephen J. Chanock, Jolanta Lissowska, Jianjun Liu, Angela Cox, Helen Cramp, Dan Connley, Sabapathy Balasubramanian, Alison M. Dunning, Mitul Shah, Amy Trentham-Dietz, Polly Newcomb, Linda Titus, Kathleen Egan, Elizabeth K. Cahoon, Preetha Rajaraman, Alice J. Sigurdson, Michele M. Doody, Pascal Gunel, Paul D. P. Pharoah, Marjanka K. Schmidt, Per Hall, Doug F. Easton, Montserrat Garcia-Closas, Roger L. Milne, and Jenny Chang-Claude. Evidence of geneenvironment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genet*, 9(3):e1003284, 03 2013. doi: 10.1371/journal.pgen.1003284. URL `http://dx.doi.org/10.1371%2Fjournal.pgen.1003284`.

P. D. Pharoah, A. C. Antoniou, D. F. Easton, and B. A. Ponder. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.*, 358(26):

2796–2803, Jun 2008.

R L Prentice and N E Breslow. Retrospective studies and failure time models. *Biometrika*, 65(1):153–158, 1978. URL http://www.jstor.org/stable/2335290.

R L Prentice, J D Kalbfleisch, A V Peterson, N Flournoy, V T Farewell, and N E Breslow. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–54, 1978. URL http://www.ncbi.nlm.nih.gov/pubmed/373811.

Thomas C Putti, Dalia M Abd El-Rehim, Emad A Rakha, Claire E Paish, Andrew H S Lee, Sarah E Pinder, and Ian O Ellis. Estrogen receptor-negative breast carcinomas: A review of morphology and immunophenotypical analysis. *Modern Pathology*, 18 (1):26–35, 2005. URL http://www.ncbi.nlm.nih.gov/pubmed/15332092.

L. C. Sakoda, E. Jorgenson, and J. S. Witte. Turning of COGS moves forward findings for hormonally mediated cancers. *Nat. Genet.*, 45(4):345–348, Apr 2013.

D. Saslow, D. Solomon, H. W. Lawson, M. Killackey, S. L. Kulasingam, J. Cain, F. A. Garcia, A. T. Moriarty, A. G. Waxman, D. C. Wilbur, N. Wentzensen, L. S. Downs, M. Spitzer, A. B. Moscicki, E. L. Franco, M. H. Stoler, M. Schiffman, P. E. Castle, E. R. Myers, M. Killackey, S. L. Kulasingam, P. Fontaine, R. S. Guido, A. Herzig, H. W. Lawson, D. R. Mody, J. Waldman, M. H. Stoler, J. M. Cain, W. Kinney, G. Birdsong, W. R. Brewster, D. Chelmow, V. J. King, R. G. Pretorius, C. M. Wheeler, B. A. Winkler, A. G. Waxman, J. J. Kim, N. Wentzensen, P. E. Castle, D. C. Wilbur, J. Cox, I. A. Eltoum, L. S. Downs, M. Spitzer, T. M. Darragh, S. E. Greening, H. K. Haefner, E. J. Mayeaux, L. Zephyrin, D. Saslow, A. B. Moscicki, K. A. Ault, M. Chevarie-Davis, E. L. Franco, M. A. Gold, W. K. Huh, D. Solomon, F. A. Garcia, A. T. Moriarty, T. J. Colgan, M. H. Einstein,

M. R. Henry, L. Massad, K. Simon, P. Gravitt, H. W. Lawson, D. Saslow, J. Cuzick, P. Gravitt, W. Kinney, E. R. Myers, K. G. Poole, M. Schiffman, D. Solomon, M. H. Stoler, D. Solomon, D. Saslow, C. J. Cohen, M. I. Edelson, F. A. Garcia, E. Holladay, W. Kinney, H. W. Lawson, K. L. Noller, E. E. Partridge, K. G. Poole, C. D. Runowicz, R. A. Smith, A. G. Waxman, E. R. Myers, D. Saslow, D. Chelmow, E. L. Franco, F. A. Garcia, A. Herzig, J. J. Kim, W. Kinney, H. W. Lawson, M. Schiffman, M. Spitzer, J. Waldman, N. Wentzensen, and D. C. Wilbur. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *CA Cancer J Clin*, 62(3):147–172, 2012.

Bryan P. Schneider, Eric P. Winer, William D. Foulkes, Judy Garber, Charles M. Perou, Andrea Richardson, George W. Sledge, and Lisa A. Carey. Triple-negative breast cancer: Risk factors to potential targets. *Clinical Cancer Research*, 14(24):8010–8018, 2008. doi: 10.1158/1078-0432.CCR-08-1208. URL `http://clincancerres.aacrjournals.org/content/14/24/8010.abstract`.

George A. F. Seber and Alan J. Lee. *Vectors of Random Variables*, pages 1–16. John Wiley & Sons, Inc., 2003. ISBN 9780471722199. doi: 10.1002/9780471722199.ch1. URL `http://dx.doi.org/10.1002/9780471722199.ch1`.

Kala Visvanathan, Rowan T Chlebowski, Patricia Hurley, Nananda F Col, Mary Ropka, Deborah Collyar, Monica Morrow, Carolyn Runowicz, Kathleen I Pritchard, Karen Hagerty, and et al. American society of clinical oncology clinical practice guideline update on the use of pharmacologic interventions including tamoxifen, raloxifene, and aromatase inhibition for breast cancer risk reduction. *Journal of Clinical Oncology*, 27(19):3235–3258,

2009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2716943&tool=pmcentrez&rendertype=abstract.

Changbao Wu. Optimal calibration estimators in survey sampling. *Biometrika*, 90(4):937–951, 2003. doi: 10.1093/biomet/90.4.937. URL http://biomet.oxfordjournals.org/content/90/4/937.abstract.

Changbao Wu and Randy R. Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):pp. 185–193, 2001. ISSN 01621459. URL http://www.jstor.org/stable/2670358.

X. R. Yang, M. E. Sherman, D. L. Rimm, J. Lissowska, L. A. Brinton, B. Peplonska, S. M. Hewitt, W. F. Anderson, N. Szeszenia-Dabrowska, A. Bardin-Mikolajczak, W. Zatonski, R. Cartun, D. Mandich, G. Rymkiewicz, M. Ligaj, S. Lukaszek, R. Kordek, and M. Garcia-Closas. Differences in risk factors for breast cancer molecular subtypes in a population-based study. *Cancer Epidemiol. Biomarkers Prev.*, 16(3):439–443, Mar 2007.

Xiaohong R Yang, Jenny Chang-Claude, Ellen L Goode, Fergus J Couch, Heli Nevanlinna, Roger L Milne, Mia Gaudet, Marjanka K Schmidt, Annegien Broeks, Angela Cox, and et al. Associations of breast cancer risk factors with tumor subtypes: A pooled analysis from the breast cancer association consortium studies. *Journal of the National Cancer Institute*, 103(3):250–263, 2011a. URL http://www.ncbi.nlm.nih.gov/pubmed/21191117.

Xiaohong R. Yang, Jenny Chang-Claude, Ellen L. Goode, Fergus J. Couch, Heli Nevanlinna, Roger L. Milne, Mia Gaudet, Marjanka K. Schmidt, Annegien Broeks, Angela Cox, Peter A. Fasching, Rebecca Hein, Amanda B. Spurdle, Fiona Blows,

Kristy Driver, Dieter Flesch-Janys, Judith Heinz, Peter Sinn, Alina Vrieling, Tuomas Heikkinen, Kristiina Aittomki, Pivi Heikkil, Carl Blomqvist, Jolanta Lissowska, Beata Peplonska, Stephen Chanock, Jonine Figueroa, Louise Brinton, Per Hall, Kamila Czene, Keith Humphreys, Hatef Darabi, Jianjun Liu, Laura J. Van t Veer, Flora E. van Leeuwen, Irene L. Andrulis, Gord Glendon, Julia A. Knight, Anna Marie Mulligan, Frances P. OMalley, Nayana Weerasooriya, Esther M. John, Matthias W. Beckmann, Arndt Hartmann, Sebastian B. Weihbrecht, David L. Wachter, Sebastian M. Jud, Christian R. Loehberg, Laura Baglietto, Dallas R. English, Graham G. Giles, Catriona A. McLean, Gianluca Severi, Diether Lambrechts, Thijs Vandorpe, Caroline Weltens, Robert Paridaens, Ann Smeets, Patrick Neven, Hans Wildiers, Xianshu Wang, Janet E. Olson, Victoria Cafourek, Zachary Fredericksen, Matthew Kosel, Celine Vachon, Helen E. Cramp, Daniel Connley, Simon S. Cross, Sabapathy P. Balasubramanian, Malcolm W. R. Reed, Thilo Drk, Michael Bremer, Andreas Meyer, Johann H. Karstens, Aysun Ay, Tjoung-Won Park-Simon, Peter Hillemanns, Jose Ignacio Arias Prez, Primitiva Menndez Rodrguez, Pilar Zamora, Javier Bentez, Yon-Dschun Ko, Hans-Peter Fischer, Ute Hamann, Beate Pesch, Thomas Brning, Christina Justenhoven, Hiltrud Brauch, Diana M. Eccles, William J. Tapper, Sue M. Gerty, Elinor J. Sawyer, Ian P. Tomlinson, Angela Jones, Michael Kerin, Nicola Miller, Niall McInerney, Hoda Anton-Culver, Argyrios Ziogas, Chen-Yang Shen, Chia-Ni Hsiung, Pei-Ei Wu, Show-Lin Yang, Jyh-Cherng Yu, Shou-Tung Chen, Giu-Cheng Hsu, Christopher A. Haiman, Brian E. Henderson, Loic Le Marchand, Laurence N. Kolonel, Annika Lindblom, Sara Margolin, Anna Jakubowska, Jan Lubin'ski, Tomasz Huzarski, Tomasz Byrski, Bohdan Grski, Jacek Gronwald, Maartje J. Hooning, Antoinette Hollestelle, Ans M. W. van den Ouweland, Agnes Jager, Mieke Kriege, Madeleine M. A. Tilanus-Linthorst, Margriet Colle, Shan Wang-Gohrke, Katri Pylks, Arja Jukkola-Vuorinen,

Kari Mononen, Mervi Grip, Pasi Hirvikoski, Robert Winqvist, Arto Mannermaa, Veli-Matti Kosma, Jaana Kauppinen, Vesa Kataja, Pivi Auvinen, Ylermi Soini, Reijo Sironen, Stig E. Bojesen, David Dynnes rsted, Diljit Kaur-Knudsen, Henrik Flyger, Brge G. Nordestgaard, Helene Holland, Georgia Chenevix-Trench, Siranoush Manoukian, Monica Barile, Paolo Radice, Susan E. Hankinson, David J. Hunter, Rulla Tamimi, Suleeporn Sangrajrang, Paul Brennan, James McKay, Fabrice Odefrey, Valerie Gaborieau, Peter Devilee, P.E.A. Huijts, RAEM. Tollenaar, C. Seynaeve, Gillian S. Dite, Carmel Apicella, John L. Hopper, Fleur Hammet, Helen Tsimiklis, Letitia D. Smith, Melissa C. Southey, Manjeet K. Humphreys, Douglas Easton, Paul Pharoah, Mark E. Sherman, and Montserrat Garcia-Closas. Associations of breast cancer risk factors with tumor subtypes: A pooled analysis from the breast cancer association consortium studies. *Journal of the National Cancer Institute*, 103(3):250–263, 2011b. doi: 10.1093/jnci/djq526. URL `http://jnci.oxfordjournals.org/content/103/3/250.abstract`.

Paige Alexandra Oliver Maas

Personal

Born in 1987 in San Jose, California.

Education

Ph.D. in Biostatistics, *expected* 2014.

Johns Hopkins School of Public Health.

Thesis: "Absolute Risk Models: Methods and Applications."

Advisors: Nilanjan Chatterjee and Mei-Cheng Wang.

B.A. in Mathematics (with minors in Biology and Psychology), 2009.

Pomona College.

Thesis: "Coral Reef Dynamics and Predation."

Advisor: Richard Elderkin.

Honors and Awards

Fellowship Achievement Award, Division of Cancer Epidemiology and Genetics, 2013.

First Place Poster Competition Winner, "Statistical Approaches to the Analysis of Audiometric Data," Research on Aging Showcase, Johns Hopkins Center on Aging and Health, 2011.

Full Scholarship to attend the Boston University Summer Institute in Biostatistics from the National Heart, Lung and Blood Institute, 2008.

Publications

Roni Falk*, Paige Maas*, Catherine Schairer, Sandra Buys, Nilanjan Chatterjee, Theresa Lee, Claudine Isaacs, Regina Ziegler. "Alcohol and risk of breast cancer in postmenopausal women: An analysis of etiologic heterogeneity by multiple tumor characteristics." *Accepted at the American Journal of Epidemiology, June 2014.*

Lin, Frank R, Paige Maas, Wade Chien, John P Carey, Luigi Ferrucci, and Roland Thorpe. "Association of Skin Color, Race/Ethnicity, and Hearing Loss Among Adults in the USA." Journal of the Association for Research in Otolaryngology. 2012 Feb; 13(1):109-17.

* co-first authors

Presentations

Paige Maas, Raymond Carroll, Nilanjan Chatterjee. "Building Risk Models with Calibrated Margins." Poster presented at the Eastern North American Region Statistical Meetings; Baltimore, MD; March 2014.

Paige Maas, Montse Garcia-Closas, Mitch Gail, BPC3 Consortium, Nilanjan Chatterjee. "Using Risk Models to Inform Cancer Prevention Efforts." Presentation at the Division of Cancer Epidemiology and Genetics Seminar; Rockville, MD; October 2013.

Paige Maas, Mitch Gail, Nilanjan Chatterjee. "Development of an Absolute Risk Model for Breast Cancer Subtypes." Poster presented at the Joint Statistical Meetings; San Diego, CA; July 2012.

Paige Maas, Roni Falk, Catherine Schairer, Sandra Buys, Nilanjan Chatterjee,

Theresa Lee, Regina Ziegler, Claudine Isaacs. "Alcohol and Breast Cancer Risk in Postmenopausal Women: the PLCO Experience." Poster presented at the 4th Annual DCEG Fellows' Training Symposium; Bethesda, MD; March 2012.

Paige Maas, Frank Lin, Ravi Varadhan. "Statistical Approaches to the Analysis of Audiometric Data." Poster presented at Statistical Research on Aging Showcase, Johns Hopkins Center on Aging and Health; Baltimore, MD; April 2011.

Research Experience

Predoctoral Fellow in the Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, 2009-2014.

Research Assistant at the Johns Hopkins Center of Aging and Health, 2010.

Research Assistant to University of Queensland PhD candidate, 2007.

Teaching

Department of Biostatistics, Johns Hopkins School of Public Health,
Teaching Assistant:

Data Analysis Workshop, Johns Hopkins Summer Institute of Epidemiology and Biostatistics, 2013.

Survival Analysis Course, Johns Hopkins Summer Institute of Epidemiology and Biostatistics, 2013.

Introductory Statistical Methods in Public Health III, 2012.

Statistical Reasoning in Public Health II, online, 2011.

Statistical Reasoning in Public Health I, online, 2011.

Master's Level Methods in Biostatistics IV, 2011.

Master's Level Methods in Biostatistics III, 2011.

Master's Level Methods in Biostatistics II, 2010.

Master's Level Methods in Biostatistics I, 2010.

Service and Leadership

Student Volunteer, ENAR Spring Meeting, Washington D.C., April 2012.

Student Representative, Prospective Student Visitors Weekend, 2010-2012.

Captain and Organizer, Johns Hopkins Biostatistics Intramural Soccer Team, 2009-2012.

Organizer, Johns Hopkins Biostatistics Student Welcome Picnic, 2011.

Coordinator, Johns Hopkins Biostatistics Student Computing Club, 2010-2011.

Organizer, Johns Hopkins Biostatistics Student Welcome Camping Trip, 2010.

Budget Committee for Campus Life and Activities, Pomona College, 2008-2009.

Computing Skills

Statistical Software: R, SAS, Stata, WinBUGS

Document Preparation: LaTeX, Microsoft Office

Other Programming Languages: Java, Octave